UNIVERSITY OF
CAMBRIDGE

Evaluating visually grounded language capabilities using microworlds

Alexander Oswald Kuhnle

Queens' College

August, 2019

This dissertation is submitted for the degree of
Doctor of Philosophy

Declaration

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. This dissertation does not exceed the prescribed limit of 60 000 words.

Alexander Oswald Kuhnle
November 2019

Abstract

Evaluating visually grounded language capabilities using microworlds

Alexander Oswald Kuhnle

Deep learning has had a transformative impact on computer vision and natural language processing. As a result, recent years have seen the introduction of more ambitious holistic understanding tasks, comprising a broad set of reasoning abilities. Datasets in this context typically act not just as application-focused benchmark, but also as basis to examine higher-level model capabilities. This thesis argues that emerging issues related to dataset quality, experimental practice and learned model behaviour are symptoms of the inappropriate use of benchmark datasets for capability-focused assessment. To address this deficiency, a new evaluation methodology is proposed here, which specifically targets in-depth investigation of model performance based on configurable data simulators. This focus on analysing system behaviour is complementary to the use of monolithic datasets as application-focused comparative benchmarks.

Visual question answering is an example of a modern holistic understanding task, unifying a range of abilities around visually grounded language understanding in a single problem statement. It has also been an early example for which some of the aforementioned issues were identified. To illustrate the new evaluation approach, this thesis introduces ShapeWorld, a diagnostic data generation framework. Its design is guided by the goal to provide a configurable and extensible testbed for the domain of visually grounded language understanding. Based on ShapeWorld data, the strengths and weaknesses of various state-of-the-art visual question answering models are analysed and compared in detail, with respect to their ability to correctly handle statements involving, for instance, spatial relations or numbers. Finally, three case studies illustrate the versatility of this approach and the ShapeWorld generation framework: an investigation of multi-task and curriculum learning, a replication of a psycholinguistic study for deep learning models, and an exploration of a new approach to assess generative tasks like image captioning.

Acknowledgements

This thesis is the result of a four-year journey, and many people have influenced, inspired, or otherwise contributed to the outcome along the way. First and foremost, I am enormously grateful to my supervisor, Ann Copestake, for providing guidance when I needed it, for granting me freedom when I could handle it, and for believing in the endeavour from its unlikely beginnings. Her mentorship has not just greatly influenced my work on this thesis, but will certainly shape my thinking beyond.

I would also like to thank Stephen Clark, Paula Buttery and Laura Rimell, for their encouragement and helpful discussions on earlier versions of the ideas underpinning this thesis. I have been fortunate to have had the chance to co-supervise Huiyuan Xie, Lars Hulstaert and Tom Sherborne, whose research was motivated by, and in turn inspired, thoughts around this thesis. Special thanks also to my examiners Lucia Specia and Simone Teufel for their insightful thoughts and helpful feedback on the thesis.

I am very grateful to the DELPH-IN community, for their hospitality on the summits I attended in 2016 and 2017, and more generally for their work on the linguistic tools that this thesis draws upon. Special thanks goes to Woodley Packard for his work on ACE, and particularly to Dan Flickinger for his sustained effort on the ERG as well as for his quick responses to the many super-specific issues I bothered him with.

Moreover, I would like to thank Raffaella Bernardi and her group, particularly Sandro Pezzelle, Ionut Sorodoc and Aurélie Herbelot, for the invitation to and welcoming hospitality during my stay at the Center for Mind/Brain Sciences in Rovereto in February 2017. This visit was supported by the European Network on Integrating Vision and Language.

This thesis would not have been possible without the financial support I received from the Qualcomm Award Premium Studentship and the EPSRC Doctoral Training Studentship. Moreover, I would like to acknowledge Queens' College for supporting me via the Munro Studentship during 2016/17, and the support I received on various occasions from the Department of Computer Science and Technology.

Last, but certainly not least, I am very grateful to my friends, in Cambridge as well as in Germany, for their distraction from this at times overwhelming undertaking, and particularly to Shachi for enduring me even when the going got tough; finally, to my parents, who laid the groundwork for this thesis through their unconditional support and invaluable mentorship in all aspects of life. Besides, without their repeated reminder to *"just get it done"*, I would probably still be writing…

Contents

Chapter 1

Introduction

How would you assess whether your image question answering system has learned to count? You would probably start off by collecting a set of images displaying multiple objects, and check how well the system is answering corresponding *"How many... ?"* questions. If you struggle to find sufficiently many images for a range of numbers and with varying types of objects, you might decide to create test images yourself, so that you have control over these details. To increase the difficulty of the task, you might then ask the system to distinguish between two successive numbers, and/or adjust the number of distractor objects. If the system is able to answer many but not all of your questions, you might want to dig deeper into its behaviour, to identify what imperfect counting patterns the system has picked up.

What you would likely not do – at least unless you are a machine learning researcher – is to scrape a large number of images from the internet, ask people to come up with questions about these images, filter out the ones starting with *"How many... ?"*, and interpret accuracy on the resulting evaluation set to indicate how well your system can count. Yet, this is roughly how research on visual question answering has started to investigate the same question in recent years. Why does this approach appear so different from what one would intuitively do?

Much of modern machine learning research practice is still shaped by the traditional principles of supervised machine learning, which originate from the goal of approximating a complex function based on a set of input-output data points. Prime examples of such applications are, for instance, object or speech recognition – tasks for which it is virtually impossible to explicitly formulate the algorithm of how to solve them, but for which one can comparatively easily obtain a large number of illustrative real-world instances. Machine learning is largely agnostic as to task and data, and postulates the following three principles for learning from data in general:

1. Since the best attempt to describe the complex task in question is to illustrate it with representative data points, systems are also best evaluated on data.
2. A system cannot be assessed using the data it was trained on, as otherwise trivial memorising without learning would solve the task perfectly.

3. A system should to be evaluated on data following the same distribution, so that the correct response can definitely be inferred based on the training data.

A consequence of these principles is the near-ubiquitous evaluation methodology of measuring and comparing task performance on a withheld test-split of a corresponding benchmark dataset. And indeed, for the type of opaque problems of traditional machine learning, benchmark-based evaluation is often seen to be both a more accurate and neutral measure of task performance than theory-laden approaches.

However, abilities like the initial example of counting objects in images do not fit into this description for two reasons: on the one hand, apart from the object recognition component, there are a range of systematic rules about counting which can be explicitly formulated, and consequently do not require data points to be illustrated; on the other hand, there are various expectations about how the ability to count should enable a system to systematically solve examples different from any it has previously seen – in fact, such generalisations are particularly challenging and thus the most interesting for evaluation. I distinguish this type of ***capability-focused evaluation*** with explicit patterns and theoretical implications, from the ***application-focused evaluation*** of tasks like speech recognition, which lack such a data-independent foundation and thus fit within the traditional machine learning principles.

With the advent of deep learning, models are perceived not just as more powerful, opaque 'end-to-end' approximators, but also as capable of genuine sophisticated understanding comparable to humans – as indicated by, for instance, the frequent usage of anthropomorphising language to describe their 'reasoning'. As a consequence, there is increasing interest in capability-focused evaluation, which explains the introduction of new benchmark datasets for holistic understanding tasks like visual question answering, language inference or reading comprehension.

This thesis is motivated by my interpretation of various problems related to dataset quality and evaluation practice, which emerged recently in the context of deep learning: that they can be largely attributed to the flawed yet dominant status of monolithic benchmark datasets to serve for capability- as well as application-focused evaluation. However, once these two orthogonal types of evaluation are distinguished and the inappropriateness of datasets for the latter acknowledged, previously unquestioned methodology choices need to be reconsidered. Are tasks like question answering, language inference or reading comprehension specific enough, or can relevant core abilities be better disentangled and isolated in separate sub-tasks? Is naturally occurring real-world data important, or do the various confounding factors it comes with in fact distract from the actual evaluation goal? Should the evaluation setup try to approximate the real-world application as closely as possible, or does it make more sense to facilitate unambiguous measurement of performance instead?

These considerations illustrate how significantly the distinction between the two evaluation goals may affect experimental practice. Instead of attempting to fix or adapt the traditional methodology so it can be used for capability-focused evaluation, my proposal in this thesis is

to complement static real-world datasets with configurable simulators for abstract data, and holistic benchmarks with what I refer to as ***unit-tests for deep learning***. The analogy to unit-testing in software engineering highlights key aspects: evaluate in isolation basic well-defined capabilities which are integral to the eventual task, leverage abstract data which covers all interesting corner cases, and strive for unambiguous passed/failed results as opposed to small incremental performance improvements. I argue that configurable data simulators provide an ideal toolbox which enables in-depth analysis of a model's decision making process and comparative assessment of its strengths and weaknesses.

The second part of this thesis centres around the ShapeWorld generation framework for visually grounded language data, as an example implementation of such a configurable data simulator. Its main motivation is to provide a testbed for visual question answering systems, however, I show how the framework can also be used to evaluate related tasks like image captioning. Subsequent experiments using ShapeWorld data present a detailed analysis of model performance including a range of novel findings, and more generally illustrate the evaluation practice of unit-testing for comparative and in-depth assessment of model behaviour. I want to emphasise this latter point, as the core contribution of my thesis is not limited to the concrete experimental findings around visual question answering, and instead provides a methodology for evaluating black box machine learning models which, I believe, will prove increasingly useful in the future.

1.1 Thesis outline

Chapter 2 substantiates the argument that emerging issues in the context of deep learning evaluation are symptoms of a mismatch between application-focused evaluation methodology for a capability-focused evaluation goal. As part of this, I review and categorise recent literature according to: first, observations that indicate a systematic mismatch between benchmark and task performance; second, the various types of dataset biases and why they question the appropriateness of 'real-world' data; third, problems with performance metrics or misleading conclusions due to invalid application; and fourth, how approaches to respond to these issues deviate, implicitly or explicitly, from the traditional evaluation practice.

Chapter 3 reconsiders fundamental aspects of evaluation methodology for machine learning with respect to the goal of evaluation as well as the purpose and nature of data. I argue that, complementary to real-world data as comparative application benchmark, abstract data fits better with the requirements of diagnostic capability evaluation. As a result of these considerations, I propose a novel evaluation framework complementary to existing practice, which centres around configurable data simulators and the concept of unit-testing for deep learning models. Moreover, I discuss the choice of visual question answering as target application task for the remainder of the thesis.

Chapter 4 introduces the ShapeWorld framework as example of a configurable data simulator for visually grounded language data. In contrast to other language data generation approaches, ShapeWorld implements a full closed-world formal semantics framework for its abstract visual domain, and employs a compositional grammar formalism which mirrors the world semantics. I argue that the principled handling of the language component of a simulator is a key differentiator to static benchmark datasets and ad hoc data generation, as it makes it possible to scale the approach and transfer it to related tasks.

Chapter 5 presents experimental results for a range of state-of-the-art visual question answering models on ShapeWorld data. Following the unit-testing principles, performance for different instance types is evaluated separately, instead of being subsumed as part of one monolithic dataset. Whereas previous work has concluded comparable overall performance of the selected models, my experiments reveal substantial differences in their ability to handle some data patterns. Narrowing down to an interesting observation regarding performance for spatial relations, I identify which architectural components actually contribute to superior performance, and which do not.

Chapter 6 reports on three projects which explore use cases for evaluation with ShapeWorld beyond comparative model analysis for visual question answering: first, investigating the effect of multi-task and curriculum learning on the learning process; second, taking inspiration from psycholinguistics to assess model behaviour similar to human behaviour; and third, leveraging closed-world abstract data for a novel approach to evaluate generative tasks like image captioning. I consider these projects to be first steps into novel directions, enabled by a data simulator framework like ShapeWorld.

Chapter 7 concludes the thesis, and highlights three higher-level aspects where I see my thesis contributing to machine learning research going forward: the quest for explainable AI, the emerging science of data generation, and the need for a data toolbox.

1.2 Key contributions

First, I provide an extensive survey and categorisation of problems and solution approaches related to deep learning evaluation.

Second, I present a novel and carefully motivated evaluation methodology for deep learning models based on configurable abstract data simulation and unit-testing principles.

Third, I introduce a configurable and extensible generation framework, ShapeWorld, for formal-semantics-style visually grounded language generation in an abstract closed-world domain.

Fourth, I conduct a detailed assessment and comparison of the abilities of a range of state-of-the-art visual question answering models to handle various instance patterns requiring, for instance, counting or relational reasoning.

Besides these four major contributions, secondary contributions include: an identification of which architectural modifications improve a VQA model's capability for simple spatial reasoning; an in-depth investigation of a VQA model with respect to multi-task and curriculum learning as well as its ability to understand the quantifier *"most"*; and a novel formal-semantics-based approach for evaluating image captioning.

1.3 Publications

First author, thesis-related content:

- Alexander Kuhnle and Ann Copestake (2017). *ShapeWorld – A new test methodology for multimodal language understanding*. arXiv: 1704.04517
- Alexander Kuhnle and Ann Copestake (June 2018). 'Deep learning evaluation using deep linguistic processing'. In: *Proceedings of the NAACL Workshop on Generalization in the Age of Deep Learning*. New Orleans, LA, USA, pp. 17–23
- Alexander Kuhnle, Huiyuan Xie and Ann Copestake (Sept. 2018). 'How Clever Is the FiLM Model, and How Clever Can it Be?' In: *Proceedings of the ECCV Workshops*. Munich, Germany, pp. 162–172
- Alexander Kuhnle and Ann Copestake (Aug. 2019a). 'The meaning of "most" for visual question answering models'. In: *Proceedings of the ACL Workshop on BlackboxNLP*. Florence, Italy
- Alexander Kuhnle and Ann Copestake (Dec. 2019b). 'What is needed for simple spatial language capabilities in VQA?'. In: *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*. Vancouver, Canada

Co-author, thesis-related content:

- Huiyuan Xie, Tom Sherborne, Alexander Kuhnle and Ann Copestake (Feb. 2020). 'Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity'. In: *Proceedings of the AAAI Workshop on Evaluating Evaluation of AI Systems*. New York, NY, USA

Co-author, no directly thesis-related content:

- Ann Copestake, Guy Emerson, Michael W. Goodman, Matic Horvat, Alexander Kuhnle and Ewa Muszyńska (May 2016). 'Resources for Building Applications with Dependency Minimal Recursion Semantics'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, pp. 1240–1247
- Yimai Fang, Haoyue Zhu, Ewa Muszyńska, Alexander Kuhnle and Simone Teufel (Dec. 2016). 'A Proposition-Based Abstractive Summariser'. In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pp. 567–578

Chapter 2

Background and motivation for a new approach to deep learning evaluation

Since its popularisation after the success of AlexNet (Krizhevsky et al., 2012), deep learning has had a transformative impact on the fields of computer vision and natural language processing. Motivated by progress on previous standard benchmarks, recent years have seen the introduction of a variety of more ambitious holistic understanding tasks, requiring a broad range of high-level reasoning abilities. However, there is increasing awareness of emerging issues related to dataset quality and evaluation practice, questioning some of the results on new tasks which appear impressive at first glance.

This chapter presents an overview and categorisation of literature on these issues, ultimately arguing for the need to focus on evaluation methodology. Before going into details, however, it is worth reiterating the traditional machine learning setup and its fundamental assumptions. Large-scale datasets play a central role for machine learning, which is, for instance, illustrated in the classic textbook *Pattern Recognition and Machine Learning* (Bishop, 2006, page 32):

> *"If data is plentiful, then one approach is simply to use some of the available data to train a range of models, [...], and then to compare them on independent data, sometimes called a validation set, and select the one having the best predictive performance. [...] it may be necessary to keep aside a third test set on which the performance of the selected model is finally evaluated."*

In other words, sufficiently big datasets serve as surrogate of a task for both training and evaluating machine learning models, as long as train and test (and validation) set are kept separate. I refer to this approach hereafter as the ***ML paradigm***. It comes with a range of implicit assumptions with respect to evaluation: (a) a single dataset can act as proxy for an underlying task; (b) performance scores on a withheld test split are sufficient to indicate the quality of a learned solution, at least comparatively; and (c) the type of generalisation necessary to master a task, as opposed to the training dataset, is defined 'negatively' as any mechanism that does not merely rely on exact memorisation.

Lacking a larger theoretical framework to justify these assumptions, their validity relies on being assessed empirically, that is, by judging the quality of evaluation results obtained following the ML paradigm. Ultimately, this is a subjective question which the field of machine learning research and its audience have to answer, but I expect the following characteristics to be unarguably desirable: (a) substantial improvements of benchmark scores over a baseline system should correspond to clear improvement of the task capability, and vice versa; (b) good performance should increase the perceived reliability of and trust in the abilities of the evaluated system; and (c) in the case of sub-optimal performance, evaluation results should ideally indicate weaknesses and lacking skills, to offer guidance for further progress. In particular in times when machine learning is successfully applied to a range of new tasks in novel fields, even previously unequivocally accepted methodology may need to be reconsidered.

I want to emphasise that my focus when referring to the ML paradigm is on evaluation, which needs to be distinguished from what Halevy et al. (2009) (and Sun et al. (2017)) famously referred to as the *"unreasonable effectiveness of data"*: that machine learning trained on large amounts of naturally occurring raw data often trumps more theory-laden approaches to solve a task. The crucial difference to my use of the term *"ML paradigm"* is twofold: on the one hand, Halevy et al. (2009) are concerned with training machine learning models, while the ML paradigm addresses evaluation; on the other hand, they emphasise the usage of task data which is naturally occurring, while the ML paradigm leverages data which can act as a useful surrogate for the task to evaluate.

In contrast to the effectiveness of data for training, I am not aware of recent papers explicitly arguing for the merits of the ML paradigm as evaluation approach. Nonetheless, apart from a few voices raising general concerns with current practices (Sculley et al., 2018; Lipton and Steinhardt, 2018; Hutson, 2018), the sentiment is definitely 'in the air' that monolithic benchmark datasets are the right way to evaluate deep learning. At the same time, there is an increasing number of papers pointing out deficiencies with specific datasets and evaluation results.

Section 2.1 discusses findings that go against the intuition of the ML paradigm: whereas models performing well on a benchmark dataset should exhibit reasonable behaviour and their abilities should translate to related/downstream tasks, this is frequently not the case. One attempt to explain (and in consequence avoid) these issues centres around involuntary biases in datasets, which are discussed and categorised in section 2.2. However, what is labelled as dataset bias can equally be interpreted as aspects where the reliance on real-world data for evaluation does not serve the intended purpose. Another explanation for misleading evaluation results is seen in inadequate performance metrics or flawed statistical comparison of performance scores, with more details in section 2.3. A range of papers, summarised in section 2.4, have been published recently which attempt to fix some of the identified evaluation issues, like dataset bias, by deliberately deviating from the ideal of the ML paradigm in some respect. Finally, section 2.5 discusses the question whether machine learning is in danger of a 'replication crisis', and

section 2.6 concludes the chapter by arguing that existing approaches to 'patch-fix' the reviewed problems do not go far enough in reconsidering the principles of evaluation methodology.

2.1 Mismatch of benchmark and task performance

If machine learning research at times appears to be solely concerned with beating benchmark scores and achieving new state of the art, this is ultimately justified by the expectation that progress on benchmarks translates into improved capabilities for the underlying tasks. This includes, on the one hand, that well-performing models exhibit reasonable perceived behaviour which encourages confidence in their abilities and, on the other hand, that the learned skills positively affect performance on related and downstream tasks. While the approach has generally been very successful – as progress in, for instance, speech recognition, machine translation or object recognition confirms – this section discusses examples where benchmark scores were found to be misleading and did, by all appearances, not correspond to actual progress. Since the equivalence of dataset and task performance is essential to the validity of the ML paradigm, such findings need to be investigated and may require us to reconsider fundamental assumptions.

Models learn weird behaviour. The result of training a machine learning model is supposed to be a system which handles instances of the underlying task appropriately. This does not necessarily imply that its behaviour has to resemble the way humans solve the task, and the types of errors a model makes may differ from what one observes with humans. However, a well-performing model is not expected to exhibit systematic trends of, for instance, relying on correlated but clearly irrelevant aspects of the input, or processing certain details consistently and obviously wrong – in short, it should not give the correct answer for the wrong reason.

A famous case in health care from the 1990s is a machine learning model which learned the counter-intuitive rule that pneumonia patients with a history of asthma have a reduced risk of dying from pneumonia, reflecting the intensive treatment in such cases (Caruana et al., 2015). Ponce et al. (2006) pointed out how vision models may rely on the image background to recognise an object, while Zhang et al. (2018b) investigated blind spots of deep convolutional networks. Sequence-to-sequence approaches to text normalisation occasionally exhibit *"silly errors"* – arbitrary confusions – when mapping digit sequences to textual number representations, according to Sproat and Jaitly (2016). In machine translation, Arthur et al. (2016) observed how neural models mistranslate low-frequency words into context-fitting but content-changing alternatives, where phrase-based machine translation rarely makes such mistakes. Furthermore, Belinkov and Bisk (2017) tested how robust machine translation models are to typos and other forms of noise, and found that not just did state-of-the-art systems fail even for moderate levels of noise, but adding synthetic noise to the training data also did not help the system to cope better with actual human typos. Williams et al. (2018a) investigated the latent tree structure

learned by neural models on a downstream task, reporting that they seem to neither learn a recognisable nor a consistent syntactic grammar formalism, while nonetheless outperforming various baselines which do not have the architectural capability to learn latent structures. Furthermore, Feng et al. (2018) found that iteratively removing words which are unimportant for the prediction from language input does not yield a 'plausible explanation' of the decision, but instead pathological instances which often consist of only 1-2 words and appear nonsensical to humans, yet nonetheless retain the original level of certainty of the model's prediction.

One of the most striking examples for 'weird' model behaviour are what became popular as *"adversarial examples"* in image classification: either insignificant changes to an image which lead the model to confidently predict a completely different class (Szegedy et al., 2014), or unrecognisable noisy images which the network nonetheless confidently assigns an object class (Nguyen et al., 2015). These findings sparked a series of investigations into the vulnerability of vision models to adversarial examples and the nature of generalisation for deep learning models, in particular the experiments of Zhang et al. (2017a), who found that modern vision models are capable of fitting mere noise remarkably well and with not substantially more difficulty than realistic data. Jia and Liang (2017) pointed out that, whereas vision models struggle with over-sensitivity to imperceptible noise, NLP systems instead largely exhibit over-stability in the face of semantics-altering modifications, which is a different kind of undesired reaction.

This highlights a few important observations: (a) deep neural networks are indeed very powerful 'function approximators'; (b) it is questionable to what degree they learn an approximation based on remarkably effective but nonetheless superficial pattern matching, as opposed to a deeper understanding of the problem and its context; and (c) the ML paradigm, while accurately measuring the superior approximation capacity, fails to provide meaningful evaluations for the sought-after deeper understanding capabilities – after all, none of the demonstrations of unexpected model behaviour above was indicated by noticeable problems with performance scores on the corresponding benchmark dataset.

Transfer/downstream failures. Good evaluation performance is not just supposed to be meaningful in comparison to other models on the same dataset, but should also be reflected on transfer or downstream tasks. The latter is particularly important for intermediate processing tasks which can only usefully be applied as part of a bigger system in real-world applications (Langley, 2011; Wagstaff, 2012). For instance, the verdict of Sproat and Jaitly (2016) when evaluation deep learning models for text normalisation was that the few but grave errors their model makes mean that it is not ready for deployment, despite overall good performance. Similar concerns were raised by Moosavi and Strube (2017), who observed that despite substantial performance improvements on the task of coreference resolution, these do not seem to be meaningful for downstream tasks, not even on similar datasets with consistent annotations. Talman and Chatzikyriakidis (2019) investigated the performance of language inference models trained on one and evaluated on

similar datasets, and found that performance levels are surprisingly benchmark-specific, even in cases of closely related datasets like SNLI (Bowman et al., 2015) and its successor MultiNLI (Williams et al., 2018b), which share most aspects of data collection methodology.

In computer vision, the problem of deteriorating transfer performance on other datasets than the one a model was trained on was noted by Yuille and Liu (2018) and Kornblith et al. (2019), although the latter found that ImageNet performance is predictive of relative accuracy on other tasks when the model has been fine-tuned. Recht et al. (2018) and Recht et al. (2019) illustrated an 'extreme' version of transfer performance drop in their investigation of generalisation capabilities for models trained on CIFAR-10 and ImageNet, two popular and long-standing image classification benchmarks. Having obtained a new test dataset by meticulously following the collection procedure of the corresponding original dataset, they showed how all models experience substantial performance decline. The magnitude of this drop was much larger than what, if it were an improvement, would generally be considered an improvement over state-of-the-art.

Reproduction is a trivial form of transfer where the experimenter tries to approximate the original experimental setup as well as possible. Fokkens et al. (2013) attempted to replicate results for measuring WordNet similarity and named entity recognition, and noted a range of typically undocumented sources of variation: data preprocessing, precise experimental setup, framework versioning, etc. Similarly, when trying to replicate results in reinforcement learning, Henderson et al. (2018) identified various reasons, from non-determinism of the benchmark environment to details specific to different codebases, for why even average performance may differ substantially between two different sets of random seeds. Such variation in performance due to seemingly minor experimental differences further highlights the questionable practice of basing improvement claims on comparatively small score differences.

Strong baselines and model arbitrariness. A more subtle case of misleading benchmark performance is caused by 'unexpectedly' strong baseline models, that is, for instance, when a slightly modified version of a supposedly weak baseline is found to perform competitively to state of the art, or when a reproduction study comparing various models indicates that none improves substantially upon the baseline level. This means, on the one hand, that the dataset is easier than previously thought and, on the other hand, that performance improvements between models were mostly meaningless and thus arbitrary. Both Ponce et al. (2006) and Sturm (2014) discussed such apparent *"performance limits"* as likely signalling that evaluation data is either insufficiently, or too challenging for better models to really make a difference. Considering the number of such cases identified for recent datasets, this hints at a more fundamental problem with the types of datasets being created for evaluation purposes, as well as with the scale of performance differences that are fallaciously considered improvements.

The following list summarises cases for different tasks where the unexpectedly strong performance of a baseline model called recent progress and previously identified state-of-the-art models into question. Cases where a simple baseline performs well by leveraging some form of dataset bias are discussed in section 2.2. Note that I do not want to imply that these systems should necessarily perform worse, but that the fact that their performance was seen as *"surprisingly"* strong suggests that expectations were not met regarding the conclusions from previous experimental results and/or the appropriateness of the corresponding datasets as evaluation benchmarks.

Computer vision: Standard image denoising architectures perform competitively with (practically) no training (Ulyanov et al., 2018).

Image captioning: Simple nearest-neighbour methods (Devlin et al., 2015) and models using a reduced bag-of-objects representation instead of the full image (Wang et al., 2018b) perform on a par with more sophisticated models.

Image synthesis: Supposedly superior GAN variants do not improve upon the original GAN model in a large-scale analysis (Lučić et al., 2018).

Language model & similar tasks: First, properly tuned, the classic LSTM cell has repeatedly been shown to perform as well or better than more recent RNN cell alternatives (Jozefowicz et al., 2015; Melis et al., 2017; Merity et al., 2018). Second, simple convolutional networks are competitive with traditional recurrent sequence models (Bai et al., 2018). Third, short-range concatenation of word embeddings performs on par with more sophisticated long-range attention mechanisms (Daniluk et al., 2017).

Lexical inference relation classification: Vanilla cosine similarity is as effective as specialised similarity measures (Levy et al., 2015).

Machine translation: A unified encoder/decoder model without attention and instantaneous output after each processed input word performs competitively with recent seq2seq models (Press and Smith, 2018).

Polysemy: Random sense assignment improves the performance of word sense embeddings as much as learned vectors (Dubossarsky et al., 2018).

Pronoun disambiguation & Winograd Schema Challenge: Simple unsupervised language models perform very well on some datasets (Trinh and Le, 2018).

Reading comprehension: A range of simple models has been shown to work well, like an entity-centric classifier (Chen et al., 2016), a single feed-forward network focusing on the last input sentence (Srinivasan et al., 2018), or a CRF tagger plus BiLSTM (Petrochuk and Zettlemoyer, 2018).

Reinforcement learning: On many benchmarks, policies parametrised by a deep neural network and trained via gradient descent are not required and/or exhibit worse performance

than technically simpler approaches, like linear- or radial-basis-function-parametrised policies (Rajeswaran et al., 2017), nearest-neighbour-based policies (Mansimov and Cho, 2018), canonical evolution strategy algorithms (Chrabaszcz et al., 2018), or random search algorithms (Mania et al., 2018).

Sentence representation learning: Simple pooling techniques perform on par with more sophisticated techniques (Shen et al., 2018), like unsupervised averaging plus common component removal (Arora et al., 2017). Moreover, both a BiLSTM architecture without any training (Conneau et al., 2018) and random sentence encoders (Wieting and Kiela, 2019) are strong baselines.

Visual Dialogue: Simple canonical correlation analysis ignoring visual and dialogue input performs competitively (Massiceti et al., 2018).

Visual question answering: Properly tuned, the classic CNN-LSTM baseline model (Lu et al., 2015) or even a simplified CNN plus bag-of-words version (Jabri et al., 2016) achieve better performance than initially reported, and can perform competitively when augmented with a simple attention mechanism (Kazemi and Elqursh, 2017). Similarly, the stacked attention network, which is often considered a baseline in the context of the CLEVR dataset, has been shown to improve markedly after tuning (Santoro et al., 2017).

2.2 Dataset bias

The quality of datasets is of fundamental importance to machine learning, since both the training and evaluation of models rely on it. Two developments in recent years changed the characteristics of datasets used in machine learning research. Deep learning improved the ability of models to handle raw real-world data and thus led to a shift towards 'end-to-end' tasks and data which, crucially, does not require complex annotation schemes and trained annotators for intermediate representations. Simultaneously, the advent of crowdsourcing platforms like Amazon Mechanical Turk made it possible to easily deploy simple annotation tasks to a pool of crowd-workers, and consequently obtain large quantities of annotated data cheaply and quickly. The synergy of both developments enabled researchers to create datasets of unprecedented size for a variety of new and more ambitious tasks. However, subsequent investigations have uncovered issues with many datasets, most of which can be summarised as different kinds of data bias.

2.2.1 What is dataset bias?

In the context of machine learning, datasets act as representatives for a specific task and hence are supposed to exhibit characteristic patterns and correlations. As ***dataset bias*** I define the coincidental systematic artefacts in the data which are not characteristic of the task in question,

even if they happen to coincide frequently[1]. Such biases can affect the learning process by suggesting heuristics which, while valid for the dataset, are invalid for the task in general. From a practical perspective, biases are undesirable as the learned behaviour may not successfully transfer to other instances of the task. More generally, the inferred behaviour is not plausible and hence not trustworthy, since it conflicts with our understanding of how the task should be solved. From the viewpoint of evaluation, dataset bias allows models to 'cheat', that is, achieve good performance results while avoiding to solve the actual task.

Before moving on to the various types of dataset bias, I will highlight how such biases can be discovered – after all, datasets are created with a genuine interest in providing a good proxy of the task. The direct approach is to identify concrete patterns reflected in the statistics of the dataset including, for instance, a skewed distribution of target classes, or correlations between the presence of certain words and the corresponding answer. Moreover, unreasonably high performance of a baseline model which is not expected to be capable of solving the task can hint at bias patterns. Another indirect approach is to destructively modify the data in a systematic way by, for instance, changing word order or replacing words. If models achieve a similar level of performance on the invalidated data, this also reveals problematic patterns. Another method for uncovering dataset bias is to investigate whether learned behaviour can be transferred to other datasets for the same task. Robust behaviour should still perform well, while heuristics relying on dataset bias are likely to fail.

2.2.2 A taxonomy of biases

As part of reviewing the recent literature around examples for dataset biases, I will introduce a categorisation of different types of biases. Many of the names are taken from the literature, however, they have not always been presented as a general type of bias, and definitely not as part of a dataset bias taxonomy as is done here. The different biases are by no means completely orthogonal, but rather useful categories to identify distinct mechanisms which may cause an observed pattern, even if actual patterns are often the effect of a combination of these mechanisms.

Selection bias. The choice of which data points are selected to represent a task may introduce unintended biases. For instance, datasets involving images often consist of photographs taken from the web. These images were created by humans who were not primarily interested in faithfully representing the visual world, but in aesthetic, social, humorous and other aspects. The ***capture bias*** encompasses preferences for point of view, position, size, lighting, occlusion, amongst others, and researchers have repeatedly warned about the problems this may imply (Pinto et al., 2008; Torralba and Efros, 2011; Tommasi et al., 2015): (a) that such data is actually

[1]Note that this definition does not address a broader version of data bias which, regardless of how 'characteristic' of the task they are, are undesirable for societal or ethical reasons.

repurposed and thus somewhat artificial; (b) that it exhibits only a limited range of patterns which may be easy to pick up but which will not generalise to the real world; and (c) that there is ultimately a danger of *"creeping overfitting"* (Torralba and Efros, 2011) to these specifics as opposed to making progress on the task. The *"Name That Dataset!"* game of Torralba and Efros (2011) probably illustrates the effect of selection bias best: a classifier which is trained to detect which of the 12 recognition datasets an image originates from, reaches an accuracy of 39% as opposed to chance level of 8%, despite the fact that all datasets are supposed to resemble the same task and thus should be hard to distinguish. Another effect noted already by Ponce et al. (2006) is that models often focus on the image background as opposed to the object to be recognised, which happens to correlate well with the desired output for the specific selection of images in some datasets.

In natural language processing, text taken from a single source is prone to the ***genre bias***. Most prominently, the Penn Treebank (Marcus et al., 1993), which dominated natural language processing research for many years, consists of articles from the Wall Street Journal, and models over time were optimised to the point where they rely on domain-specific patterns and experienced significant performance drops on non-newspaper text (Rimell and Clark, 2008; Manning, 2011). On the one hand, researchers are aware of this type of bias and sometimes actively try to diversify datasets (Wang et al., 2018a); on the other hand, however, there is a tendency to accept progress on one dataset as progress for a general ability to understand language.

Another instance of selection bias can be summarised as ***annotation scheme bias***. Annotated datasets typically impose a discrete set of 'non-natural' labels on their data points, based on strong assumptions about the world. As such, an annotation scheme introduces undesired bias (not considered characteristic of the task) in cases where it forces to choose a category when either multiple labels seem equally suitable or none is applicable. For instance, Tommasi et al. (2015) referred to ***category/label bias*** as patterns due to poorly defined semantic categories, and ***negative bias*** as effects caused by the finite set of distinct categories on the 'rest of the world', both of which are clearly caused by the choice of object classification scheme. This type of bias was particularly common in natural language processing before the deep learning era, when different syntactic and semantic annotation schemes were central to most datasets. For instance, Manning (2011) noted how the remaining performance gap for part-of-speech tagging on the Penn Treebank may be mostly due to inconsistencies in the annotation – or spurious 'consistency' in ambiguous cases – and not a question of actually improving the tagging ability. More abstract and/or semantic annotations struggle even more with label and negative bias, that is, to clearly distinguish their categories and specify their scope. The semantics of prepositional phrases, for instance, has seen multiple annotation scheme refinements (Litkowski and Hargraves, 2006; Srikumar and Roth, 2013; Schneider et al., 2015), but also faces the difficulty of separating cases of 'regular/productive' prepositional phrases from other usages (Baldwin et al., 2009).

While the annotation scheme bias in natural language processing was reduced with the move towards end-to-end tasks, recent datasets at the same time became more prone to the ***annotator bias***, which comprises patterns caused by annotators. Contrary to skilled annotators (often the researchers themselves) and elaborate annotation schemes, crowdsourcing involves many more confounding factors which can result in unintended patterns. Smith (2012) raised the problem of annotator bias by describing the task represented by annotated datasets as *"what a particular set of annotators, with a particular kind of training, on a particular kind of data, within a particular amount of time, would generate on the test set"* (Smith, 2012). This concern applies in particular to the crowdsourcing setup, where the choice of annotators, their education and the annotation conditions is far less rigid. Perhaps surprisingly, this concern is only rarely expressed: Gururangan et al. (2018) conjectured that crowd-workers develop annotation strategies, and that the framing of the annotation task can have a significant effect to the point of priming certain such strategies, as they convincingly show to be the case for the SNLI Dataset (Bowman et al., 2015); and Petrochuk and Zettlemoyer (2018) identified ambiguity problems in the SimpleQuestions reading comprehension benchmark (Bordes et al., 2015) caused by the annotation process.

Input relevance bias. Many tasks involve multiple inputs and, in addition to processing each of them separately, require the ability to combine their information. Input relevance bias refers to situations where supposedly relevant information is not necessary to achieve good performance. Most obviously this is the case for multimodal tasks like visual question answering, where a question is supposed to be answered based on an accompanying image. ***Modality bias*** refers to the systematic tendency that one modality suffices to infer the correct output with high confidence. Multiple examples were reported for the VQA Dataset (Antol et al., 2015) indicating this type of bias. Zhang et al. (2016) noted how a language-only model which completely ignores the image can answer almost half of the questions correctly, among these 78% of the binary yes-no questions. Subsequently, Agrawal et al. (2016) observed how seemingly well-performing models jump to conclusions after only the first few question words, thus concluding that they fail at complete question and image understanding. Goyal et al. (2017) pointed out that indeed the first two to three words of a question represent a strong prior for the correct answer. Adding to this, Mudrakarta et al. (2018) achieved 44.3% accuracy for empty questions, above 50% when only keeping the word *"colour"*, and even more when preserving other 'most attributed' (that is, most important for a model's decision) words like *"many"*, *"what"*, *"how"*. Cirik et al. (2018) noted another case of modality bias for the visual referring expression dataset Google-Ref (Mao et al., 2016), where good performance can be achieved even after discarding the referring expression entirely. Thomason et al. (2019) identified modality bias in the EmbodiedQA and Interactive Question Answering Dataset, which makes it possible for trivial baselines to outperform many other systems. Similarly, Anand et al. (2018) showed that biases in the EmbodiedQA dataset allow their blindfold baseline which ignores the visual input to achieve state-of-the-art performance.

Related effects have been observed for natural language inference. Here, a system is presented with two input sentences, a premise and a hypothesis, and is required to analyse their logical relation, that is, whether one entails the other, they contradict each other or are neutral with respect to entailment. The ***hypothesis bias*** corresponds to the effect of the hypothesis input alone being sufficient to infer this relation. Gururangan et al. (2018) have shown that this is the case for the SNLI and MultiNLI datasets (Bowman et al., 2015; Williams et al., 2018b). Poliak et al. (2018) extended the investigation to ten language inference datasets and found that in six cases a hypothesis-only model outperforms a majority-class baseline. From this observation, they conclude that the hypothesis-only model is a more appropriate indicator of the lower performance bar (or dataset bias) for a language inference dataset than the commonly used majority-class baseline, given that the task is supposed to involve both premise and hypothesis. Furthermore, Levy et al. (2015) showed a version of hypothesis bias for the task of classifying the lexical inference relation between two words, where it suffices to identify whether one of the words is a prototypical hypernym.

Another example of input relevance bias can be found for reading comprehension, where a question based on a text passage needs to be answered, and ***question/passage bias*** means that a system only relying on one of the two inputs can be comparatively successful. Kaushik and Lipton (2018) analysed five reading comprehension datasets for both these effects, including the CBT (Hill et al., 2016), the CNN/DailyMail (Hermann et al., 2015) and the SQuAD dataset (Rajpurkar et al., 2016), by randomising in each case the passage-question assignment.

Wang et al. (2018b) have illustrated how a generalised version of this bias type, ***input component bias***, can be investigated for image captioning – technically a single-input task – by decomposing the image input and considering various approximate representations based on location/size/centrality/etc of objects in the image. In their experiments, Wang et al. (2018b) found that image captioning models for MS-COCO (Lin et al., 2014) can learn to produce reasonable captions merely by knowing about the objects in an image while ignoring, for instance, their location and relation.

Data statistics bias. A complex dataset allows for a range of different perspectives to obtain summarising statistics of its data points. A simple representation for a classification dataset, for instance, is the instance distribution over the set of categories. Data statistics bias corresponds to the situation where a certain perspective reveals that a dataset exhibits unfavourable trends which affect evaluation quality. What exactly is seen as 'unfavourable' depends on the focus: for instance, Horn and Perona (2017) argued that image classification datasets suffer from ***uniformity bias***, as the distribution of object classes is chosen uniformly in contrast to the distribution of objects encountered in the real world. For evaluation, however, uniform distributions are generally preferred as they avoid trivially successful majority-label responses.

More frequently considered an issue with datasets is ***simplicity bias***, when the data actually follows the *"long tailed"* real world distribution (Horn and Perona, 2017) and is thus dominated by comparatively simple patterns. In natural language processing, this phenomenon is well-known as Zipf's Law for various kinds of dataset statistics. Kafle and Kanan (2017b) analysed the VQA Dataset (Antol et al., 2015) and observed that improving accuracy for *"is/are"* questions by 15% increases overall performance by over 5%, whereas answering all *"why/where"* questions correctly corresponds to only 4.1%, clearly illustrating a case of simplicity bias. Linzen et al. (2016) noted how an effect of such bias is that models learn flawed heuristics for syntactic dependencies on a prediction task, which fail on harder instances. In a thorough theoretical analysis of the bAbI dataset (Weston et al., 2015), Lee et al. (2016) showed how 18 of the 20 sub-tasks can be seen as variations of the same containee-container relationship setup and consequently are not actually that 'different'. Trichelair et al. (2018) analysed the Winograd Schema Challenge (Levesque et al., 2012) and found that, despite the explicit requirement to not contain associativity and predictable structure, many such patterns can be found in the dataset.

Overstability/-sensitivity bias. The terms *"over-stability"* and *"over-sensitivity"* were introduced by Jia and Liang (2017) and describe models which are overly robust to semantically meaningful, or overly sensitive to semantically meaningless modifications of the input. While these terms, on the surface, clearly describe system behaviour, they can be interpreted as symptoms for a category of dataset biases which encourage to learn such behaviour, instead of the intended correct solution.

A common example of over-stability bias is the tendency of models to be insensitive to word replacements within a related word class, thus referred to as ***word class bias***. Sproat and Jaitly (2016), for instance, noted that sequence-to-sequence models for text normalisation sometimes arbitrarily confuse number terms. Mudrakarta et al. (2018) found that trained visual question answering models do not change their answer for questions with content words replaced by either semantically related hyponyms or arbitrary nonsensical words of the same part-of-speech. As a prominent sub-type of word class bias, ***hyponym bias*** was also investigated by Shekhar et al. (2017) in their *"foiled"* version of the MS-COCO (Lin et al., 2014) image captioning dataset. They found that, based on the predictions of systems trained for image captioning by inferring the most likely output, it is often not possible to identify erroneous captions where one noun was replaced by a hyponym.

Overly stable or sensitive behaviour is also encouraged by the fact that for many datasets, despite superficial complexity, noting the presence of some key words suffices to infer the correct response, thus referred to as ***signal word bias***. Mudrakarta et al. (2018) investigated this effect as the *"attribution"* of words, by which they refer to the influence of a word on a model's decision. They show how, in the case of visual question answering, replacing low-attributed but important phrases does not affect the answer, and for the task of answering questions about spreadsheets,

dropping stop words has a significant effect. Moreover, for both question answering tasks and for reading comprehension, they analyse model behaviour when adding pre-/suffix words to the language input. Their findings show that the *"attribution"* of words in the pre-/suffix is a strong indicator of whether the modification will have an effect on a model's decision or not. Dasgupta et al. (2018) experimented with sentence embeddings for natural language inference and found that negation words as well as antonyms (of a word in the premise) in the hypothesis is used as signal for contradicting sentence pairs, while word overlap acts as a signal for entailment. Another example of signal word bias in the context of visual referring expressions was indicated by Cirik et al. (2018), who found that shuffling the words in a referring expression or dropping all words but nouns and adjectives does not affect the recognition ability of models on the Google-Ref (Mao et al., 2016) dataset. They conclude that models largely base their decision on the tendency of one word being a *"prototypical hypernym"* in isolation, that is, whether it *"tends to be entailed"* (vs *"tends to entail"*).

The experiments of Mudrakarta et al. (2018) were partially motivated by another type of over-sensitivity bias: that models for reading comprehension tend to rely on the last words/sentences in the input, and thus exhibit ***input recency bias***. They investigated various adversarial methods of adding a misleading sentence to the reading comprehension paragraph, some of them chosen based on the specific input or the evaluated model's sensitivity/gradients, and observed decreasing accuracy throughout. Similarly, Srinivasan et al. (2018) found that focusing on the last sentence of the input context improves model performance on a story cloze test.

2.3 Performance metrics

Besides evaluation data, performance metrics are the second fundamental pillar of machine learning evaluation methodology. Metrics try to approximately assess the quality of performance of a system without the need to conduct human studies to judge its behaviour, or deploy a model in the intended downstream application to observe its impact. Requirements consequently include: metrics should be automatic, fast to compute, and ideally summarise performance in a single benchmark score to facilitate comparison with other models. The following paragraphs review the problem of trying to capture the evaluation of generative models in single-number metrics, and the statistical fallacies and inadequacies of comparing performance scores.

Performance metrics for generative models[2]. Discriminative models are relatively straightforward to assess, as there are well-understood metrics like accuracy or precision/recall which

[2]Here and in the following, *"generative model"* refers to the informal usage of the term in the context of deep learning as a model which generates new data (images, language, etc), as opposed to a *"discriminative model"* which produces classification labels from a fixed set of categories. In traditional machine learning, these terms are more narrowly formally defined: considering a function $f\colon x \to y$ to be learned (e.g., a classifier), a generative model is a model of the joint probability distribution $p(x, y)$ of inputs and outputs, whereas a discriminative model is a model of the conditional probability distribution $p(y \mid x)$ of outputs given inputs.

are easy to interpret. In contrast, the output space for generative tasks is high-dimensional and the expected response is not well-defined, meaning that there is no single correct output. As a consequence, it is unclear how to quantify distances between potential outputs in a meaningful way, and what best characterises the quality and appropriateness of 'good' outputs. Theis et al. (2016) illustrated that three common metrics for generative image algorithms like GANs are largely independent for high-dimensional data, and Lučić et al. (2018) noted how a *"memory GAN"* just reproducing the training data would score perfectly in most current evaluations. Barratt and Sharma (2018) identified a range of problems with the recently introduced Inception score, both with the metric itself and with its popular adoption by the vision community, emphasising the need for *"meaningful evaluation metrics"* over *"ad-hoc metrics"* (Barratt and Sharma, 2018). Xu et al. (2018) assessed a range of common metrics for GANs for desirable characteristics like distinguishing generated from real images, sensitivity to mode dropping/collapsing, or detecting overfitting, and found that popular metrics did not overall cover these aspects well.

In the case of image captioning, Anderson et al. (2016) noted how metrics are primarily sensitive to n-gram overlap with gold captions, which is neither necessary nor sufficient for improving human judgement of generated captions. Low correlation between captioning metrics and human judgement was already identified as a problem by Elliott and Keller (2014). Kilickaya et al. (2017) investigated the robustness/sensitivity of various metrics to synonyms, word order, phrase replacement and similar modifications, and found that semantically close captions may receive differing scores whereas captions with different meaning but surface similarity do not. Similar concerns about the correlation with human judgement were expressed and experimentally confirmed by Liu et al. (2016) for the evaluation of dialogue systems, and by Sulem et al. (2018) for text simplification. A possible reason for this mismatch is seen in the fact that metrics like BLEU, METEOR or ROUGE originated from machine translation and were adopted for respective task, despite differences in what qualifies a good solution. However, Callison-Burch et al. (2006) noted early on that even machine translation is overly reliant on BLEU despite performance increases being neither necessary nor sufficient for improved translation quality, and pointed out use cases where BLEU should and should not be used for evaluation. Furthermore, Post (2018) highlighted problems with changing parametrisation and reference processing schemes, resulting in substantially different performance scores, while Reiter (2018) reviewed reports of (non-)correlation of BLEU with human evaluation for language output quality assessment and concluded that its use outside of machine translation is questionable.

Originally used for speech recognition evaluation, perplexity is another common performance metric for generative prediction models. However, Smith (2012) pointed out problems with this measure: on the one hand, it unnecessarily requires the model to be probabilistic and the comparability of scores is highly sensitive to details of the event space; on the other hand, improved perplexity scores are known to not correlate well with actual error reduction in application tasks (Chang et al., 2009; Smith, 2012).

Statistical flaws of interpreting performance scores. While the performance metrics for discriminative tasks itself are well-defined, concerns have been raised repeatedly about statistically sound comparison between scores and what can be concluded from them. Ioannidis (2005) famously summarised the problems around likelihood of statistically significant false-positive findings in the context of systematic biases, unreported failure results and simultaneous experimentation by multiple research teams. Bennett et al. (2009) presented a striking example of a deliberately nonsensical experiment which investigated whether a dead salmon can correctly determine emotional state when shown photographs of humans. According to standard statistical analysis, the high-dimensional fMRI scans imply significantly positive results, however, the absurd conclusion highlights how standard statistical thresholds are ineffective in controlling for multiple comparisons. Considering this problem in the context of machine learning, Demšar (2008) pointed out that the ease of generating new algorithms thanks to flexible machine learning frameworks, in combination with the practice of relying on significantly improved benchmark scores, implicitly encouraged many such false-positive findings. Arguably, deep neural networks and frameworks like TensorFlow and PyTorch nowadays allow for even more architecture variation than ten years ago.

An interesting theoretical analysis by Szucs and Ioannidis (2017) concluded that null hypothesis testing is unsuitable for large datasets, since increasing the sample size guarantees that the null hypothesis can be rejected eventually even with miniature effect sizes. Reimers and Gurevych (2018) demonstrated a related effect by comparing an architecture with itself (BiLSTM-CRF architecture on seven common NLP sequence tagging tasks). If the test is based on predictions of two trained versions of this model, they found significant differences more frequently than what the 5% level of $p = 0.05$ would suggest. However, when comparing the *"learning approach"* with itself – that is, the performance score distribution of multiple training runs, which takes into account the various sources of randomness in modern ML training – the relative amount of significantly different results is as expected at the 5% level. They concluded that the common approach of assessing significance based on a single run is problematic, and that randomness in modern ML can have substantial effects on model comparison.

These concerns focus on problems with statistical comparison methodology when applied properly. However, a recent review of papers published at the conference on Neural Information Processing Systems in 2017 (Király et al., 2018) assessed the mere completeness of argumentative steps, and found substantial shortcomings for most papers: besides missing baseline scores as reference, only around a third of the papers reported confidence intervals, however, with no reference or explanation, and only 3% reported formal comparison/hypothesis testing.

2.4 Approaches to fix evaluation

There have been approaches to address the problems discussed in the last sections and fix the broken evaluation methodology. I distinguish between approaches which see the problem on the side of the dataset not appropriately reflecting the task and its difficulties, thus trying to fix the data, and ones which consider the precise task formulation and evaluation setup as less suited for informative evaluation, thus improving the task. Overall, these examples indicate that there is awareness of more precisely what aspects of evaluation need changing, and what ideal conditions would look like.

2.4.1 Fix data distribution

Improve datasets. A dataset is supposed to represent the task to be assessed. One reason for why an evaluation does not yield the desired insights can consequently be identified in the quality or appropriateness of a dataset. An issue may be that the original train-test split of a dataset is considered too 'unspecific', in which case this split can be fixed. For instance, Atzmon et al. (2016) created a split of the MS-COCO captioning dataset (Lin et al., 2014) requiring compositional generalisation[3], and Agrawal et al. (2017) introduced C-VQA, a compositional split of the VQA Dataset (Antol et al., 2015). Agrawal et al. (2018) released another split of the VQA Dataset with changing priors, called VQA-CP, where test answer distributions per question type differ from training distributions, and Li et al. (2018) presented ZST-VQA which requires zero-shot transfer to succeed on the test split. A similar attempt is to fix the data distribution by extending the dataset: Zhang et al. (2016) balanced yes-no questions of the VQA Dataset's subset of abstract instances, whereas Goyal et al. (2017) introduced VQA 2.0 in which each question is associated with a pair of similar images that result in different answers. In a similar spirit, Linzen et al. (2016) proposed to evaluate models on naturally occurring sentences sampled based on their grammatical complexity, to counter the bias towards simple constructions in natural data, which is detrimental to the quality of evaluation when focusing on syntax-sensitive dependencies.

A more drastic measure is to improve the dataset from the ground up, by introducing a new version which addresses the problems found for the predecessor. For instance, MultiNLI (Williams et al., 2018b) is the successor of SNLI (Bowman et al., 2015) and improves its genre diversity, or SQuAD 2.0 (Rajpurkar et al., 2018) succeeds SQuAD 1.0 (Rajpurkar et al., 2016) and introduces unanswerable questions. Other examples include the visual question answering dataset of Kafle and Kanan (2017a) which contains various balanced question categories including absurd questions and more appropriate evaluation metrics, both targeting shortcomings of the VQA Dataset. The NLVR dataset (Suhr et al., 2017), besides consisting of synthetic data, explicitly

[3]Compositional generalisation refers to the ability to understand the meaning of a new phrase like *"red square"*, given understanding the concept of a *"square"*, the colour *"red"*, and the fact that a *"square"* can be coloured.

aims to improve upon the VQA Dataset with respect to the crowdsourcing task setup and thus avoid resulting biases. Zellers et al. (2018) introduced the SWAG dataset and the methodology of adversarial filtering[4] in combination with oversampling to construct more challenging datasets semi-automatically, while avoiding biases due to direct crowdsourced data collection. However, Torralba and Efros (2011) have discussed this phenomenon of introducing new datasets in response to weaknesses of a 'predecessor' dataset using the example of the historical sequence of image classification datasets. Importantly, they argued that such a process is likely doomed to result in a *"vicious cycle"* of ad hoc improvements, unless one reconsiders the underlying mechanisms which cause undesired dataset bias.

Challenge set. Besides generically fixing an existing dataset, unsatisfying evaluation results can also be addressed by carefully collecting a set of particularly challenging instances. This relates to the proposal of Linzen et al. (2016) to sample instances based on complexity, since it explicitly acknowledges and bypasses the problem that arbitrary real-world data is dominated by simplicity and under-represents the difficult phenomena for which one would like to analyse model behaviour (Isabelle et al., 2017; Ettinger et al., 2017; Wang et al., 2018a). The FraCaS test suite for textual inference problems (Cooper et al., 1996) is an early example of such a challenge set. More recent examples include the Winograd Schema Challenge (Levesque et al., 2012) for world knowledge and common-sense reasoning; a challenge set for neural machine translation (Isabelle et al., 2017); the *"Build it break it"* workshop and shared task (Ettinger et al., 2017) in which models are adversarially tested by users (see also Smith (2012)); the ARC AI2 Reasoning Challenge for question answering (Clark et al., 2018) which consists of natural science questions taken from standardised tests and even includes a further challenge subset; the hand-crafted diagnostic test suite as part of the GLUE benchmark (Wang et al., 2018a); or the corpus of precise natural textual entailment problems introduced by Bernardy and Chatzikyriakidis (2018) which is presented as a successor of the FraCaS suite addressing various shortcomings (Chatzikyriakidis et al., 2017).

Artificial/abstract data. Fundamentally, real-world data may not offer the degree of control over the data distribution which is required to test specific behaviour, in particular systematic generalisation (Mitchell et al., 2018). In contrast, artificial data makes it possible to control minutiae details of the data, and thus has often been leveraged to implement in-depth analyses of model capabilities: for instance, the ability to handle patterns generated by various formal grammars (Gers and Schmidhuber, 2001; Avcu et al., 2017; Suzgun et al., 2019); dialogue simulation (Scheffler and Young, 2001; Scheffler and Young, 2002); logical reasoning (Bowman, 2013; Evans et al., 2018); contextual language command understanding (Dukes, 2014; Bisk

[4] Adversarial filtering refers to the process of filtering data with the aim to affect model performance negatively, rather than according to model/experiment-neutral criteria – that is, focusing on instances which are often referred to as *"adversarial examples"* in the context of deep learning.

et al., 2016); image scene semantics and visual saliency (Zitnick et al., 2016); the bAbI tasks for reading comprehension and question answering (Weston et al., 2015); correct identification of long-distance dependencies (Linzen et al., 2016); simple compositional generalisation in visual question answering (Johnson et al., 2017a); IQ tests for neural networks (Hoshen and Werman, 2017; Barrett et al., 2018); translation of complex language commands into sequences of actions (Lake and Baroni, 2018; Bastings et al., 2018); the quality of image generation with more appropriate metrics (Lučić et al., 2018); subitising and approximate numerosity (Wu et al., 2018); lookup table composition as a combination of memorisation and compositional retrieval (Liska et al., 2018); robustness of reinforcement learning (Zhang et al., 2018a); symbol rewriting abilities (Weber et al., 2018); visual-relational same-different problems (Kim et al., 2018); understanding of recursive syntactic structure and compositionality (Paperno, 2018); or TextWorlds for relational reasoning in reading comprehension (Labutov et al., 2018).

Evaluation according to the ML paradigm should be based on a randomly selected subset of the data, so that test data follows the same distribution and is thus, in principle, learnable from the training data. However, this reasoning is increasingly challenged, particularly in the context of natural language processing. For instance, Weber et al. (2018) talk about *"linguistic generalisation"* going beyond the typical meaning of *"generalisation"* in machine learning. Lake and Baroni (2018) pointed out the fact that the vast majority of sentences are unique even in huge corpora, highlighting the 'counter-intuitive' generalisation capabilities of sentence-based models to largely unseen instances. Marcus (2018) introduced the useful distinction of ***interpolation*** versus ***extrapolation***, that is, the ability to generalise to similar inputs on the one, and to unseen novel inputs on the other hand. Artificial data is particularly useful to evaluate this latter type of extrapolating generalisation, since it requires the train/test data to follow a specific structure which is hard to achieve with for real-world data.

The ability to control the content and difficulty of individual instances is often mentioned as an advantage of using artificial data (Avcu et al., 2017; Evans et al., 2018; Zhang et al., 2018a). Moreover, abstract data reduces noise, ambiguities and reliance on common sense or world knowledge which are considered irrelevant for the evaluation goal (Zitnick et al., 2016; Johnson et al., 2017a). Language in an abstract domain may even result in more interesting patterns for these reasons (Bisk et al., 2016). Generally, Weston et al. (2015) aptly summarised the role of evaluation using artificial data: *"While any learner that can solve these tasks is not necessarily close to full reasoning, if a learner fails on any of our tasks then there are likely real-world tasks that it will fail on too (...)."* (Weston et al., 2015).

2.4.2 Fix evaluation task

Minimal pairs. Some problems are broad and unspecific in the sense that they require a range of abilities to arrive at the correct response, but that the task formulation does not explicitly try to identify sub-optimal behaviour. As a consequence, the setup offers opportunities to 'cheat' the

evaluation and achieve good performance by leveraging superficial indicators in the data instead of the intended inference mechanism. To counter this effect, test instances can be designed as minimal pairs, where a second distractor instance is chosen or modified to be deceptively similar but wrong, and the task is rephrased as a binary distinction task. This evaluation methodology is superior in at least two respects: on the one hand, performance on the resulting balanced binary classification task is measured as accuracy which is the most straightforward metric and, on the other hand, the way distractors are chosen makes it possible to target specific abilities which are required to distinguish the two instances.

The Winograd Schema Challenge (Levesque et al., 2012; Levesque, 2014) is a prime example of such a task. Its instances consist of sentences with a carefully designed binary pronoun/possessive reference ambiguity, which can be resolved if the situation described by the sentence is interpreted correctly, and which is difficult to resolve otherwise since simply swapping two key words would result in the opposite answer. Hodosh et al. (2013) argued that image description understanding is better framed not as a generative task via image captioning, but as a ranking task where the model needs to be able to tell which of two (or more) descriptions is more appropriate for a given image. Similarly, Stanovsky and Hopkins (2018) proposed Odd-Man-Out puzzles as a flexible task format which, on the one hand, makes otherwise opaque concepts like word similarity concrete via minimal-pair-style evaluation while, on the other hand, is nonetheless intuitive enough to obtain annotations via crowdsourcing.

A simple approach is to modify existing datasets to create a minimal pair version, ideally via an automatic way of transforming data point into decoy instances. Zhang et al. (2016) introduced an extended balanced binary version of the VQA Dataset (Antol et al., 2015) focusing on yes/no questions, so that questions alone have no inherent answer bias. Chao et al. (2018) generalise this principle to *"question- and image-only unresolvable"* instances for multiple-choice visual question answering. Furthermore, Mahendru et al. (2017) combined VQA instances with an additional image for which exactly one of the question premises is false. In the context of image captioning, Hodosh and Hockenmaier (2016) automatically swapped, replaced, added or removed phrases in correct captions to obtain distractor captions, while Shekhar et al. (2017) focused on a more specific analysis by replacing a single noun per caption with hyponyms based on MS-COCO super-categories (Lin et al., 2014).

Different to such linguistically targeted modification, Ding et al. (2016) chose similar decoy captions from the dataset based on paragraph vector similarity. Glockner et al. (2018) modified instances from the SNLI dataset by a single word, via synonyms and hypernyms for entailing instances, and exclusive co-hyponyms and antonyms for contradicting instances. Marvin and Linzen (2018) automatically constructed pairs of English sentences with one being ungrammatical, and evaluated whether language models attribute a higher probability to the grammatical sentence. Rosenfeld et al. (2018) analysed the unstable behaviour of object detection models on modified images where regions are transplanted with known object regions from other images.

Trichelair et al. (2018) switched the two candidates in sentences from the Winograd Schema Challenge wherever possible and, in addition to assessing performance on the modified dataset, they also checked consistency between the original and modified instances.

Probing. One problematic aspect of deep neural networks is generally considered to be their black box nature, which makes it hard to reason about what the strengths and weaknesses of models are. While task-focused evaluation indicates how well a problem is solved, it does not analyse model performance in more detail. A recently increasingly popular approach is to follow a different evaluation methodology based on probing (Dasgupta et al., 2018; Conneau et al., 2018; Cirik et al., 2018) – also called *"stress-testing"* (Naik et al., 2018; Geiger et al., 2018) – network behaviour with the aim to shed light on specific aspects of what a model has learned.

A technique particularly useful to assess the information contained in learned representation embeddings is to use them as basis to solve an auxiliary predictive task. For instance, Adi et al. (2017) analysed sentence embeddings for how well they enable a classifier to predict sentence length, word order and whether a word was part of a sentence. Conneau et al. (2018) extended this idea to a suite of ten tasks around surface information (sentence length, word content), syntactic information (bigram shift, tree depth, top constituent) and semantic information (tense, subject/object number, semantic odd-man-out, coordination inversion).

Another common approach is to test the sensitivity of the model output to modifications of the input. As already mentioned in section 2.2, Jia and Liang (2017) characterised two problematic reactions one may observe: over-sensitivity to semantically meaningless noise, or over-stability to semantics-altering changes. They probed reading comprehension systems with various methods of adding a distracting sentence to the paragraph in question. Dasgupta et al. (2018) sampled comparison sentences from the SNLI dataset (Bowman et al., 2015) and modified word order, swap comparison words, or introduce negation, to test the sensitivity of a model when judging comparison-based inference patterns. Similarly, Cirik et al. (2018) probed the handling of visual referring expressions when shuffling or dropping words, to the point of discarding the referring expression entirely, and Mudrakarta et al. (2018) analysed more systematically which words can be dropped for question answering systems without affecting their answer decision.

Naik et al. (2018) introduced a stress-test suite for natural language inference covering three classes of tests, targeting competences like antonymy or numerical reasoning, robustness to distractions like high word overlap or negation words, and various forms of noise. In computer vision, Hendrycks and Dietterich (2018) presented the ImageNet-C variant with 15 common visual corruptions and Icons-50, which assess robustness of a model to visual corruption and surface variation, respectively. Geirhos et al. (2018) applied parametric distortions to a 16-class version of ImageNet and compared model with human performance when the signal gets weaker. Equivalently, Sturm (2014) proposed analysing the robustness of music information retrieval systems to audio transformations which do not alter characteristics relevant for the retrieval task.

Focusing on more specific linguistically informed phenomena, Jumelet and Hupkes (2018) devised a set of tasks to assess whether language models can handle negative polarity item constructions. Geiger et al. (2018) assessed the ability of language inference models to correctly understand sentences involving multiple quantifiers in combination with modifiers and negation. Finally, Goldberg (2019) investigated the syntactic abilities of the BERT model (Devlin et al., 2019) to correctly identify subject-verb agreement on a variety of inputs.

Inspiration from psychology. An interesting approach to evaluate machine learning, which recently gained some attention, is to implement experiments and apply experimental methodology from psychology. Given that researchers increasingly describe their models using 'anthropomorphising' attributes like *"understanding"*, *"attention"* or *"remembering/forgetting"*, it makes sense to investigate their behaviour similar to how human behaviour is analysed. Ritter et al. (2017) emphasised the promising value of the rich heritage of cognitive psychology for better understanding deep learning models, and advocate a hypothesis-driven approach where input-dependent behaviour predictions are confirmed or refuted in specifically designed experiments. In their study, they replicated a well-established experiment to investigate shape bias when learning to associate objects with word labels. Re-assessing the shape vs texture hypothesis, Geirhos et al. (2019) found that ImageNet-trained CNNs behave differently from humans, but by adapting the data can learn more robust human-like shape-based representations.

Another example are the experiments of Nematzadeh et al. (2018) around theory of mind, who constructed scenarios described by natural language of non-trivial belief relations between various people about the location of objects, together with questions of increasing difficulty addressing either reality or first- and second-order beliefs about object locations. Similarly, Eysenbach et al. (2016) assessed whether neural network models can recognise at what point in the course of a short visual story which person holds incorrect beliefs about the state of the world. Drawing inspiration from psycholinguistics, Mhasawade et al. (2018) analysed the learnability of non-/conservative quantifiers[5], in parallel to experiments showing that children are only able to learn new determiners corresponding to conservative quantifiers.

2.5 Are most ML research findings false?

The provocative title of this section deliberately alludes to the seminal paper of Ioannidis (2005). Mainly in response to a replication crisis in medical research, this paper identifies high-level mechanisms which lead to a systematic increase in 'false' findings, that is, claimed effects which are subsequently refuted. The observations around evaluation practice can be transferred to other

[5]Conservative quantifiers implicitly restrict the quantification-relevant sets to the quantified noun. For instance, *"Half the squares are red."* is equivalent to *"Half the squares are red squares."*, since other *"red"* objects are not relevant to the statement's interpretation.

scientific fields, and I think it is worth considering the following six corollaries postulated by Ioannidis (2005) in the context of deep learning research in recent years:

1. *"The smaller the studies conducted in a scientific field, the less likely the research findings are to be true."*
2. *"The smaller the effect sizes in a scientific field, the less likely the research findings are to be true."*
3. *"The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true."*
4. *"The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true."*
5. *"The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true."*
6. *"The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true."*

A few researchers have expressed their concern about research practice in machine learning related to these points: Sculley et al. (2018) noted a lack of *"empirical rigour"*, Lipton and Steinhardt (2018) commented on *"troubling trends in ML scholarship"*, Hutson (2018) recently even voiced the question that 'hangs in the air': is machine learning facing a replication crisis?

Following a literal interpretation of *"replication"*, one may respond that machine learning is well guarded against such a crisis, given the fact that experiments can easily be repeated, particularly thanks to the increasingly common practice to release paper-accompanying code and data online[6]. I propose to look at *"replication"* from a different angle and thereby, I believe, capture the concerns about machine learning practice more faithfully: the type of replication crisis ML research may be facing is not due to an inability to reproduce the experiment, that is, the performance number of a model on a specific dataset, but to reproduce the implied/promised superior capabilities of this same model, which the ML paradigm implies. Ioannidis (2005) linked the amount of such spurious improvements to the *"prevailing net bias"* in the community. Indeed, continued experimental practice despite the range of findings reviewed in this chapter, which report weird model behaviour, transfer/downstream failure, dataset biases and inadequate performance metrics, can only be attributed to a strong belief in the abilities of deep learning.

[6] However, Hutson (2018) rightfully pointed out that: (a) the majority of papers still do not come with open-sourced code; (b) experiments are sensitive to minuscule aspects of the training conditions down to random seeds (Henderson et al., 2018) and hardware details; (c) the same level of significance may not be replicable due to flawed statistical methods (Szucs and Ioannidis, 2017; Reimers and Gurevych, 2018; Király et al., 2018); and (d) the scale of experiments coming from research groups in industry is simply unfeasible to replicate for academic researchers.

This belief is further testified by the common usage of anthropomorphising and (deliberately?) imprecise language like models learning to *"understand"*, *"infer"*, *"attend"*, *"recognise"*, as opposed to more technical terms related to optimisation, to describe model behaviour (Levesque (2014) and Lipton and Steinhardt (2018) mention the problem of language as well). Anthropomorphising language may have the effect of sustaining this belief in overall progress of the field, and at times fool even more cautious researchers into over-optimism despite doubts about a range of individual experimental results.

What is the reason for *"prevailing net bias"* to be able to cause a replication crisis? Belief in the potential for human-like abilities of deep neural networks lowers the threshold of willingness to accept results suggesting such capabilities. Consider the not infrequent situation where a qualitative analysis based on a few data points reveals both positive and, crucially, negative evidence – *"here the model fails to. . . "* – but it is nonetheless concluded that the model performs better thanks to its superior capabilities, as confirmed by a few percent improvement on a benchmark. Do we really expect that the superior ability in question would improve performance by only, say, 1-3%? Instead of accepting the hypothesis of the model being superior, it should probably be questioned (referring back to the assumptions underlying the ML paradigm from the introduction of this chapter): (a) whether the dataset really is a good surrogate for the evaluated task; (b) whether an improved performance score is sufficient to support the claim of superior capabilities; and (c) whether the test set even requires the respective abilities to be solved.

The problematic dominant role of benchmarks for evaluation is referred to by Ioannidis (2005), as that *"the high rate of nonreplication [...] is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study"*. While benchmarks start off as a useful tool for comparative evaluation of different approaches to solve the same task, over time research focuses solely around them as the dominating factor for acceptance of results within the community (Sculley et al., 2018). By standardising training data and evaluation procedure, attention shifts primarily to creation and evaluation of new models for a task or, in other words, *"machine learning for machine learning's sake"* (Wagstaff, 2012), or *"mindless comparisons among the performance of algorithms"* (Langley, 2011).

Why *"mindless"*? First, a single performance score provides very limited insights into the relative strengths and weaknesses of a model and, as a consequence, offers little guidance for the most impactful focus of future research (Langley, 2011; Sculley et al., 2018). Second, taking a dataset as the desired objective does not indefinitely reflect and challenge the interesting core abilities of the underlying task in a progressing field (Pinto et al., 2008; Torralba and Efros, 2011; Wagstaff, 2012). Third, comparatively little attention is paid to translating progress on a dataset into corresponding improvements on the real-world application that inspired the benchmark in the first place (Wagstaff, 2012; Chiticariu et al., 2013; Sturm, 2014). In the worst case, systems with improved application performance for certain instance types are not recognised due to the fact that overall dataset performance is not much affected.

Taken together, these aspects indicate the lack of a 'regulariser' for the process of introducing new models – that is, similar to the machine learning technique, a mechanism which keeps the model development process in balance by, for instance, requiring a certain degree of robustness, generalisation and transferability, to prevent unhindered community-wide benchmark overfitting. Consequences of the latter can definitely be observed: a large number of task-specific holistic architectures with wildly varying names and seemingly arbitrary variations of all parts of a system (see, for instance, the multitude of visual question answering models mentioned in 3.3), as opposed to generic network modules whose beneficial effect is uncontroversial, like batch normalisation (Ioffe and Szegedy, 2015) or residual connections (He et al., 2016). Moreover, the importance of hyperparameter search instead of robust learning processes, or the perception of model building as *"dark art"* versus the existence of a rich set of proven best practices.

As Torralba and Efros (2011) argued, better benchmark datasets are unlikely to let us escape the *"vicious cycle"* of dataset creation. To overcome the detrimental effect of monolithic benchmarks, interest has to shift from cheap model comparisons as the driving force, to detailed evidencing of model capabilities where benchmark scores play only a minor role as comparative *"sanity checks"*. Sturm (2014) identifies the lack of control over the content of evaluation data which a benchmark dataset can possibly offer as the fundamental problem, and illustrates this point vividly with the example of *"Clever Hans"*, a horse which supposedly exhibited extensive arithmetical, reasoning and language understanding skills (Pfungst and Rahn, 1911). Driven by scepticism towards the hypothesis of a 'clever' horse, a sequence of experiments testing alternative explanations for the apparent evidence under carefully controlled conditions eventually yielded a far more likely explanation of the observations: subconscious, nearly undetectable micro-cues given by the person posing the question to the horse.

2.6 Conclusion

The central proposition of this thesis is that evaluation methodology in deep learning can be more flexible than the currently dominant practice of following the ML paradigm suggests. As reviewed in section 2.4, researchers have experimented with a range of approaches deviating either implicitly or deliberately from the ML paradigm to obtain more interesting evaluation results. However, in my opinion few if any of them go far enough in reconsidering the principles of evaluation methodology – which I will do in section 3.1 – and instead try to 'patch-fix' the issues within the traditional framework. Nonetheless, these approaches indicate that there is awareness of more precisely what ideal evaluation conditions would look like, if they were possible to be realised within the existing methodology. For instance, taking the example of dataset bias, researchers largely agree why certain imbalances are undesirable for an appropriate evaluation of the qualities of a system, but it is less obvious how to obtain unbiased real-world datasets.

In contrast, I view these issues as, in the bigger picture, being less problems in themselves, but rather symptoms of a more fundamental flaw: that the ML paradigm and its assumptions around monolithic benchmark datasets are increasingly insufficient as basis to evaluate the kinds of broader, more natural tasks and higher-level capabilities researchers of 'artificial intelligence' are becoming interested in. What is needed instead is a different evaluation framework which lets one incorporate the knowledge and hypotheses one has about a task. In the following chapter, I will illustrate how, beyond stating one's expected improvements and observing whether they happen to translate to (minor) improvements in benchmark performance, these expectations can be 'implemented' by specifically designing experiments and data to clearly convince others of – or refute – the superior capabilities of a proposed model.

Chapter 3

Evaluation methodology: underlying taxonomy and a proposal

Machine learning is an experimental field: deep neural networks are best treated as 'black boxes', and the only good way to analyse them is via empirical assessment. Evaluation methodology is thus of paramount importance to the field, and I have argued in chapter 2 that recent practice is unsatisfying and prone to wrong conclusions. The central contribution of this chapter is a proposal for a novel evaluation approach which offers more detailed insights into model behaviour. Section 3.1 discusses data- and evaluation-related concepts and distinctions which constitute a useful framework for thinking about evaluation, confirming that the ML paradigm is not suited for detailed model investigation. Based on these considerations, section 3.2 describes the proposed methodology in detail and justifies design choices. Finally, section 3.3 introduces the task of visual question answering, and argues that it is an interesting choice to illustrate my evaluation methodology.

3.1 Taxonomy of methodology-related aspects

Given the various problems with current machine learning practice, it is helpful to take a step back and reconsider fundamental assumptions related to evaluation methodology and the ML paradigm. In the following, I discuss the goal of evaluation (section 3.1.1) as well as the purpose (section 3.1.2) and nature (section 3.1.3) of data. In particular, I introduce terminology and motivate distinctions which constitute a framework for thinking about approaches to deep learning evaluation. The ML paradigm is one such approach, the unit-testing-inspired approach I propose as part of this thesis is another.

3.1.1 Evaluation goal

It is important to be clear about the goal of a specific experimental evaluation of a machine learning model. On the one hand, the main purpose may be pragmatic and focused primarily on solving a practical problem or improving a real-world system. On the other hand, the intention may be to fundamentally analyse and improve a model's capabilities to handle certain input patterns. While the aim of the latter ultimately is to feed into the former, the two are best understood as separate goals. The choice of evaluation goal influences many aspects of the evaluation setup, from the characteristics of task and data, to performance metrics, to what constitutes meaningful improvements.

Pragmatic and application-focused evaluation. Machine learning is an optimisation/automation technique and thus used to improve existing applications or make it possible to efficiently tackle new problems. Pragmatic evaluation consequently focuses on an existing practical application and the data that can be observed/collected in its context. Moreover, it is often obvious what the application should be optimised for, which either corresponds directly to a suitable performance metric like accuracy or precision/recall, or can be measured when plugged into the wider system, for instance, via A/B-testing. In other cases, the ideal performance metric needs to be approximated, which makes the results less reliable and should thus be handled more carefully. A popular example of an approximate metric is BLEU for machine translation and its problems (see section 2.3).

The data for a real-world application is generally complex, domain-specific, reliant on context, and thus often exhibits spurious correlations. This makes it hard to disentangle the various core abilities required to solve the task and obtain a clear signal for what, if anything, has led to improved overall performance. As a consequence, static benchmark datasets are prone to overfitting when used repeatedly, which is why progress should regularly be checked against 'downstream tasks' and/or on new data from a live system.

Fundamental and capability-focused evaluation. A natural way to make progress on a task is to consider abilities which are likely to be helpful. These may be lower-level concepts inspired by optimisation theory or statistics, or higher-level concepts inspired by how, according to our understanding, humans approach and solve a problem. It is important to clearly specify the investigated capability, as this in turn defines what ideal evaluation data should look like. Assessing abstract concepts does not necessarily require natural data or a realistic use case, and in fact it is often easier to illustrate an ability with abstract data which focuses on the appropriate details and eliminates other confounding factors. For instance, taking the example of counting, a real-world evaluation setting is neither necessary nor does it strengthen experimental results in evidencing the (abstract) ability to count. Moreover, the evaluation task and performance metric are ideally chosen to facilitate unambiguous experiments, like accuracy for a balanced set of

minimal pair instances (see section 2.4.2). The expected result is a binary indicator of whether a model passed or failed the evaluation – has learned the ability or not – and thus is expected to either reach 100% accuracy or remain well below.

The more artificial the experimental setup, the more there is the danger of improvements being meaningless in real-world applications due to flaws in the evaluation. Usually, such spurious improvements can be attributed to the fact that the investigated capability was insufficiently specified and the experiments thus ended up testing for unintended simpler abilities. A reason may be poor or insufficiently controlled data quality leading to, for instance, dataset bias which makes it possible to 'cheat' the test, or relying on deceptively intuitive but vague capability descriptions, like *"multimodal inference"* or *"intelligent behaviour"*, for which it is unclear how to assess them. The history of the Turing Test famously illustrates the difficulty of designing a convincing test for the not well-understood concept of *"intelligence"* (however, Turing himself introduced the test mainly as a pragmatic way to avoid fundamental discussions about such elusive concepts).

Note that, in principle, overfitting is not a problem here, but rather the desired result: a model that manages to 'overfit' a suitable capability-focused benchmark can be said to have learned the required ability to solve it. Another difference from application-focused evaluation is that higher-level abilities can often be broken down into more specific components, thus enabling iterative progress.

Current practice: superposing pragmatic and fundamental practice. Many of the recently introduced tasks, like visual question answering, focus on a vaguely specified human-inspired capability, like multimodal reasoning, instead of a practical application. As a consequence, datasets have to be artificially created, although crowdsourcing evokes the impression of the resulting data being natural and application-focused. Based on this appearance, it is justified to follow the practice of relying on ever so slight performance improvements as an indicative signal for incremental progress on a real-world task, despite the fact that such semi-artificial data (see section 3.1.3) is particularly prone to unintended biases and confounding correlations. Whereas capability-focused evaluation should aim to reduce the possibilities of overfitting to patterns other than the investigated ability, following this practice may turn out to be particularly detrimental to the quality and persuasiveness of experimental results.

3.1.2 Purpose of data

Which function data is supposed to fulfil within machine learning determines ideal characteristics or, conversely, when introducing new data it is important to take its intended purpose into account. The following distinctions present key aspects to consider:

- ***Fixed vs flexible source***: Presentation of and interaction with the data can follow either a more rigid or a more variable design, including its availability.

- ***Application-driven vs hypothesis-driven structure***: The higher-level structure of the data may be chosen with either a general task or a specific testable hypothesis in mind.
- ***Generic vs model-informed content***: The content of the data may either make little to no assumptions or be tied to details of the model class or even instance.

In the following paragraphs, three common functions of data within machine learning are discussed with respect to these points: as training data, as comparative benchmark, and for in-depth evaluation.

Training. The most obvious use for data is to train machine learning models. Deep learning is comparatively insensitive to data quality but definitely profits from vast quantities of data points. It is further common practice to augment data in various ways: for instance, by increasing the frequency of underrepresented classes, applying semantics-invariant transformations, augmenting with auxiliary tasks on the same data, leveraging related sources, to name a few. All this clearly suggests a preference for a flexible data source. Moreover, its structure is largely determined by the application that the trained model is supposed to solve. Finally, while the content of training data is generally expected to be generic, to enable training of any type of model, model-dependent augmentation like the addition of adversarial examples is not uncommon.

Comparative benchmarks. The relative qualities of machine learning models are usually assessed by comparing performance on benchmark data. Fair comparison requires standardisation of the evaluation procedure, consequently a fixed data source is preferred here. This includes presenting data as a single dataset with accompanying evaluation script, and eliminating model-unrelated confounding aspects which may affect results. In addition, such a dataset is ideally 'temporally fixed', that is, used over the years to facilitate comparison to older models, and with 'fixed access', that is, limiting repeatedly running evaluations to tune a model which ultimately leads to (community-wide) overfitting, for instance, by restricting the number of submissions for evaluation on a withheld test dataset. Besides overfitting, controlling access is also important to ensure the statistical validity of results given the problem of controlling for multiple comparisons, which otherwise are likely to yield some positive results just by chance. Benchmarks moreover require the structure to be driven solely by the problem in question, to preserve comparability despite changing hypotheses, which would otherwise intrinsically favour certain methods. Similarly, its content needs to be agnostic to any modelling aspects and cannot rely on properties like being probabilistic or using neural network techniques.

In-depth evaluation. Data constitutes the only viable approach to investigate most deep learning models, which are otherwise hard to interpret. Detailed – as opposed to comparative – evaluation benefits from a flexible data source which facilitates controlling and adapting all aspects of the data. Since models may have different strengths and weaknesses, the data needs to

be flexible enough to enable meaningful analysis in either case. Additionally, the structure of evaluation data is usually driven by changing hypotheses about what aspect of model behaviour is considered most interesting to assess more thoroughly. For instance, one may focus on relational reasoning as a suspected weakness and thus require data containing relational instances which challenge this capability. In particular, structure here needs not necessarily resemble an underlying task, but may test the limitations of a model on unrealistic and/or adversarial instances. Since a model usually introduces new techniques and architecture design decisions to address shortcomings of previous models, the content of data for in-depth evaluation is expected to be informed by these aspects to obtain the most convincing results. This may go as far as using details like model outputs or gradients to design adversarial evaluation instances.

Current practice: monolithic datasets. Following the ML paradigm, the majority of recent work centres around monolithic datasets which serve as training data as well as benchmark plus, in some cases, the basis for more detailed evaluation. The latter, however, is limited by what additional annotations a dataset provides like, for instance, a more fine-grained categorisation of instance types, which makes it possible to report performance per category. Otherwise, the practice of qualitative evaluation by hand-picking a few illustrative examples is a questionable way to infer properties of a model. By concentrating on a single monolithic dataset, each of the aforementioned purposes of data suffers: (a) training: many different sources of training data could be utilised instead of just one; (b) comparative benchmark: the quality of benchmarking is affected due to lacking 'fixed access' and consequent overfitting; and (c) in-depth evaluation: analyses are severely limited by the annotations a dataset provides, while not at all informed by what motivated design decisions for the analysed model.

3.1.3 Nature of data

Data comes in different forms, like numbers, text, images, etc. However, to be useful for current machine learning, a large set of equally-shaped data points is required. The shape is what unites the variety of instances for a task and thus characterises the space of valid data points: for instance, image captioning examples fundamentally consist of an image and an accompanying textual caption, possibly further constrained to be a single sentence, whereas a simple form of language inference may consist of two sentences – premise and hypothesis – plus an associated inference label indicating entailment/neutrality/contradiction. Not every point of this shape is valid – for example, two sentences cannot be entailing and contradicting at once – but is subject to various constraints, like logical consistency in the previous example. Exactly what constitutes a valid example of a real-world problem can be hard to specify formally, to the point of defaulting to an *"I know it when I see it"* specification. Importantly though, a more concrete characterisation of the underlying structure of data provides a framework for how to reason about it. For instance, syntactic theory identifies useful components like words or phrases and rules like

valid phrase compositions, which inform the processing and analysis of natural language data. Finally, some instances may be more common than others, which is captured by the distribution over the space, specifying the relative frequency of valid data points. An explicit space structure helps to formulate this distribution in a more meaningful way, like the distribution over words instead of over (mostly unique) sentences.

The nature of data is a defining property in this context, with ***natural real-world data*** on one end versus ***artificial abstract data*** on the other. Besides these two 'extremes', I introduce a third category which is referred to as ***semi-artificial data*** and which, I argue, aptly describes many of the labelled and/or crowdsourced datasets used in modern deep learning. While this data appears to be natural at first glance, this name emphasises the fact that it has, crucially, several artificial features.

Natural real-world data. It is hard to pin down exactly what makes data 'natural', which is why it is often intuitively defined as *"like/from the real world"*. While this vaguely describes the natural data space, it lacks specificity with regards to the structure and thus its distribution can only really be characterised as given by random samples. This generally results in extremely sparse coverage, but some superficial structure can help to densely approximate its characterisation. Note that real-world distributions often resemble a power law distribution, that is, the distribution is dominated by a few patterns and exhibits a long tail of relevant but rare points. For instance, most written sentences are unique, but their distribution can be roughly captured when interpreted in terms of components like words, n-grams or syntactic/semantic representations, all of which resemble power laws. With respect to evaluation, it is thus no surprise that natural data, lacking explicit structure, comes in the form of a large number of randomly sampled data points, and that the 'best-possible' differentiation of train and test data is based on a simple random split.

Artificial abstract data. I consider data 'artificial' if it is either created with a specific problem in mind or transferred from its natural context to the problem in question. Synthetic abstract data is the prime example of fully artificial data, while semi-artificial data is discussed below. The structure of the artificial data space is known by design and explicitly specified by its generating mechanism, which defines its rules and meaningful components. The distribution can thus be controlled in detail – although global patterns emerging from the interaction of different components can still be difficult to predict. The component distribution is usually chosen to be uniform, as there is no 'natural' reason to differentiate frequencies for abstract content. In contrast to natural data, artificial data is ideally represented by its generating process and not by a fixed dataset. On the one hand, a generator makes it possible to create datasets of any size and configuration when required. On the other hand, not all structural aspects may be obvious by looking at data points of a fixed dataset, but are explicit in the generator specification. The more

a generator supports the configuration of its parameters, the less its application is constrained to one specific task, but its data can be useful for a variety of problems. As a consequence, training and test data are not required to follow the same distribution.

Current practice: semi-artificial data. One purpose of the distinction between natural and artificial data is to highlight how many of the recent labelled and crowdsourced datasets are best described as 'semi-artificial', as opposed to fully natural. On the one hand, labelled datasets usually introduce an artificial discrete classification which is supposed to uniquely characterise any instance. Depending on the application, these classes are more or less obviously chosen, but even in seemingly straightforward cases like object recognition or parsing, existing categorisations are controversial (Tommasi et al., 2015; Manning, 2011). On the other hand, crowdsourced datasets are collected by posing an artificial task to human workers precisely because such data does not naturally occur. While platforms like Amazon Mechanical Turk are expected to lead to more 'natural' annotators than, for instance, university students or subject experts (Smith, 2012), many other aspects of the crowdsourcing setup are artificial, like people doing crowdsourcing as a paid job with a consequent bias to solve tasks quickly and simply (Gururangan et al., 2018). If the dataset includes images, these are often sourced from available photo datasets – like MS-COCO based on Flickr (Lin et al., 2014) – which show staged scenes selected by human photographers based on aesthetic, social, humorous and other criteria (Pinto et al., 2008).

The degree of artificiality depends on the dataset and can be controlled to some degree by the data collection methodology. So what is the key difference to natural data? The more guided/enforced collection process implicitly shapes the nature of data which, while still being opaque, cannot anymore be characterised as *"like the real world"*. In particular, it may introduce non-natural biases and artefacts which are not intentional, but simultaneously hard to avoid or detect, given the opaque structure of natural data in the first place. The implication is that while such data in many ways approximates natural data well, we cannot rely on the fact that it does so in every respect, and consequently have to question its status as proxy for a real-world application. There is a danger that instead of combining the advantages of being natural as well as task-focused, such data in fact ends up being 'opaquely artificial' and thus irrelevant (as some examples in chapter 2 illustrate).

3.2 Unit-testing for deep learning

I have argued that currently dominant evaluation practice following the ML paradigm is suited for comparative benchmarking (if task and data are appropriate), but does not offer the right setup for more detailed model investigation. In response to this situation, I propose an approach to evaluation largely orthogonal to the ML paradigm, which I refer to as ***unit-testing for deep***

learning. Using the terminology from section 3.1, unit-testing leverages abstract data to support in-depth and capability-focused evaluation. Section 3.2.1 expands on the methodology and the definition of unit-tests, and section 3.2.2 justifies some of the practical design choices. The remainder of the thesis then centres around a concrete implementation of this approach for visually grounded language capabilities: the ShapeWorld data generation framework.

3.2.1 The unit-testing evaluation methodology

Instead of aspiring to use natural data, unit-testing embraces the characteristics coming with artificial data. The underlying data space is chosen as an abstract ***microworld*** which is well-suited to illustrate the range of capabilities one is interested in evaluating. Importantly, its definition explicitly comprises the fundamental mechanics and semantics of the space like, for instance, rules of causality or spatio-temporal im-/possibilities. The microworld is abstract in the sense that its structure is chosen not primarily to reflect realistic conditions, but to accurately cover the concepts under consideration, and thus its relevance mainly lies within the specific evaluation context. For instance, counting objects may be seen as an abstract capability which is independent of what type of object is being counted, in which case data to assess the counting proficiency of a model does not require realistic scenes but can be simplified to arbitrary abstract objects, to focus on the 'essence' of generic counting instead of conflating it with real-world object recognition. However, such data will not on its own constitute a good application-focused benchmark to evaluate the performance of, for instance, a system which is supposed to estimate the number of people in a crowd.

Central to the unit-testing approach is a ***configurable data simulator*** which generates microworld instances representing concrete evaluation cases. Their properties and distribution can be controlled in detail by the highly structured data space of the parametrised simulator. Within the evaluation framework provided by such a simulator, a ***unit-test*** corresponds to a complete configuration of the generator engine, that is, a specification of what data patterns should be produced, with the aim to analyse a certain aspect of model behaviour. The unit-testing methodology thus frames experiment design as the process of 'designing data', guided by a concrete hypothesis. This is in contrast to the established practice of using existing, comparatively generic (that is, not hypothesis-driven) benchmark datasets – a practice which is rarely discussed and justified in more detail.

However, the unit-testing proposal is not just concerned with designing data, but also with guiding the focus of experiments. Capability-focused evaluation in machine learning is, I argue, necessarily an iterative and falsification-driven process. It is ***falsification-driven*** in that the null-hypothesis should be that a model does not exhibit a certain ability, and it is the obligation of experiments to convince us otherwise. It is ***iterative*** since hypotheses usually start off with a crude formulation and are refined in the course of obtaining results. These characteristics further emphasise the importance of designing data for machine learning and, as a consequence,

the necessity of a configurable data simulator framework. Note that this is in stark contrast to the ***positivism-driven*** evaluation often found in recent deep learning research, where small performance improvements, potentially backed by qualitative assessment of a few selected instances (sometimes even including failure cases), are supposed to evidence the assumption that a model has indeed learned a certain ability (see also discussion in section 2.5). These are weak indicators at best, and scepticism should decide in favour of the null-hypothesis: that model performance, despite small improvements, is unlikely to be due to the acquisition of a new capability.

3.2.2 Justification of design decisions

The unit-test analogy. The proposed methodology resembles the concept of unit-testing in software development in many respects. Both evaluate well-defined functional tasks requiring a specific ability to solve them. While these abilities are ultimately supposed to be employed as part of a real-world application, unit-tests analyse their performance individually, isolated from the wider context. Unit-testing is applied to comparatively basic and abstract mechanisms, and their coverage is increased additively by expanding the number of tests, however, the aim is not to achieve application-level guarantees. In contrast, application-level tests do not try to identify and disentangle the various functions involved, and their coverage can only be reduced subtractively by artificially constraining the full task. Moreover, unit-tests try to cover the full range of potential inputs, even when they are (almost) never encountered as actual inputs in practice. Unit-test evaluation strives to produce clear binary results – passed or failed, or 100% vs chance-level accuracy – as opposed to application-level performance metrics which tend to be more complex, may consist of multiple conflicting signals, and involve various confounding factors. Similar to software engineering, both evaluations are important and largely orthogonal to each other with respect to their intended purpose.

The analogy between the two concepts is not perfect, though. A difference is that there are often no obvious atomic ‘units’ when evaluating understanding capabilities. On the contrary, a test concerned with, for instance, counting can be further specialised to focus only on certain spatial arrangements of objects or on a varying number of distracting objects, while it can simultaneously be generalised to a test covering arithmetic abilities which involves counting as an implicit subtask. In particular the fact that the assessment of model behaviour can almost always be broken down into more specific patterns is important here, as this is one of the main differentiators compared to benchmarks focusing on, for instance, the counting ability in general. Related to that, the process of introducing unit-tests here is decidedly iterative and informed by the results of previously run tests, instead of being defined independently at once, as is the case for test-driven software development. Nonetheless, despite some differences, the analogy to the software engineering concept of unit-testing gives the right intuition of how to think about the role of the proposed methodology.

Why abstract? Fundamentally, my proposal is a response to the insufficiency of existing real-world and/or crowdsourced datasets to provide the basis for in-depth evaluation of deep learning model capabilities, and the consequent lack of such analyses. Whereas the full cognitive process to solve a task is often unclear, we can nonetheless identify general core competencies that likely feed into this process. Unit-tests make it possible to investigate these in isolation by implementing and evaluating hypotheses around model behaviour. Formulating abstract tests has two advantages: on the one hand, it corresponds to the fact that the analysed capabilities are 'abstract', and thus best illustrated abstractly, that is, independent of a concrete pragmatic task (consider again the example of counting); on the other hand, natural data is too complex, opaque and noisy to provide the basis to implement such tests and obtain clear indicative results. Even crowdsourced data collected with the explicit aim to provide capability-focused evaluation has repeatedly been shown to be dominated by unintended biases (see section 2.2), and is often too inflexible to provide more detailed insights anyway – all of which can be trivially avoided by artificial data.

Data distribution and generalisation. According to the ML paradigm, data points for both training and testing are considered to be samples from a single data distribution. There are two practical reasons for this assumption: on the one hand, statistical optimisation guarantees are much harder – if at all possible – to formulate and prove if distributions differ; on the other hand, the distribution of natural data itself is not a well-defined object, which makes it hard to differentiate, whereas a more concrete characterisation would facilitate both experimentally realising and mathematically handling distributional shift. The type of generalisation required to achieve good performance within the same distribution is referred to as ***interpolation***, and it was rightfully pointed out recently that, ultimately, a more powerful type of generalisation, referred to as ***extrapolation*** or ***zero-shot learning***, is desired for many abilities (Marcus, 2018; Mitchell et al., 2018) – that is, a degree of robustness in the face of differing data distributions or even different task setups.

Data simulator engines provide the means to control the distribution to a degree far beyond what is possible with natural data, and consequently are the optimal basis to design unit-tests to evaluate extrapolating generalisation capabilities. Moreover, designing such data suggests a useful framework to think about the difference between the two types of generalisation: while assessing interpolation does not require to change the simulator configuration, but just to generate new samples from the same unit-test specification, evaluating extrapolation alters the distribution of some components within the microworld space. To use an example from ShapeWorld, two component distributions could be the shape and colour of an object, which are configured to not produce some combinations, like *"red squares"*. However, the simulator still produces *"red shapes"* and *"coloured squares"*, so the component concepts could be learned, and thus a successful model should, in principle, be able to infer the correct response even for the unseen

combination of a *"red square"*. This illustrates how the type of extrapolating generalisation test I propose materialises as a simple combinatorial constraint on a more abstract level of the structured data space, and the model is required to 'overfit' not to the lower abstraction level of combinations, but to the level of components.

On the one hand, one may debate whether all generalisation is 'compositional' in this sense – for instance, considering the human ability to handle numbers one has never come across before, or more abstract inferences like the non-existence of a biggest number. On the other hand, it could be questioned whether extrapolation is even necessary to solve real-world needs. While I do think that compositional extrapolation is necessary, I only want to present some pragmatic considerations in support of it being an interesting evaluation target:

- Efficiency: compositionality is a sample- and storage-efficient way to represent the combinatorial complexity of the real world.
- Robustness: compositional understanding effectively increases robustness to at least some types of adversarial examples which rely on the complexity of the input space.
- Interpretability: since there is clear evidence for compositionality in human cognition, models resembling this characteristic are more intuitively interpretable.

Explicit configuration. A parametrised simulator is necessary both to enable the specification of a variety of unit-tests as well as to support generalisation tests as discussed above. On top of that, the configuration parameters serve as a good way to formalise the purpose of evaluation, that is, what tasks it is supposed to address, similar to the *"datasheets for datasets"* proposal of Gebru et al. (2018). On the one hand, developers of a simulator are encouraged to identify what specific abilities the data is able to target and, consequently, what aspects of testing data one would be interested in controlling to this end. On the other hand, when using the data for evaluation, explicit configurability can inspire more thorough investigations by suggesting available parameters, and prevent misuse of data for analyses it is not designed for. Finally, while the design of data(sets) rarely starts off without any issues and misconceptions, limitations can be countered by extending the configurability and thus the expressivity of a framework, which may lead to longer-standing test suites, in contrast to datasets which are quickly superseded due to minor shortcomings (Torralba and Efros, 2011).

Progress without hardware. The unit-testing methodology separates capability-focused in-depth evaluation from application-focused benchmarking, arguing that each is best pursued individually and comes with largely orthogonal requirements. In particular, benchmarks ultimately involve having to work with huge real-world datasets, combining various technical improvements into a single model, and tuning the entire setup for optimal performance. In the context of deep learning, all this requires an increasing amount of computing time, advanced

hardware, and expertise in coordinating large-scale distributed experiments, which renders such work unfeasible for many university research groups. Moreover, concerns about ecological implications of this trend in machine learning have been expressed recently (Strubell et al., 2019). However, these demands do not exist for capability-focused experiments. Unit-test evaluation is targeted and cheap, freeing researchers who are interested in in-depth model analysis from costly benchmarks, and consequently introduces a kind of 'division of labour' where both interest groups can meaningfully contribute to the larger goal of progressing machine learning research.

Interpretability and psychology. Deep learning techniques with their non-convex non-linear optimisation are notoriously hard to analyse via mathematical proofs and guarantees. I expect that this situation will not change substantially, so alternative means of *"opening the black box of deep neural networks"* (Shwartz-Ziv and Tishby, 2017) need to be considered. One promising approach is to be inspired by empirical methodology from psychology, which is concerned with opening the 'black box of human behaviour' (see also section 2.4.2). Indeed, there are many parallels between the proposed unit-testing methodology for deep learning on the one hand, and the design and refinement of experiments for analysing human behaviour on the other. Unit-tests address concerns about the reliability of a model's decisions by thoroughly probing key assumptions about its decision mechanism, which may be the most convincing evidence for reliable performance given the issue of (non-)interpretability.

3.3 Why visual question answering?

In general, the evaluation principles introduced in the last section are applicable to a wide range of capabilities, to the degree that desirable systematic patterns can be illustrated in abstract scenarios. I believe that many of the recently popular tasks in deep learning and natural language processing research would profit from this approach – and in parts already have, to the degree that the examples in section 2.4 share aspects with my proposal. In this thesis, with ShapeWorld as a concrete implementation of a data simulator framework, I chose to focus on visual question answering (VQA, see figure 3.1 for an example), and visually grounded language understanding more generally, as a typical task sharing many characteristics with other recently popular tasks: (a) it is a broad task comprising diverse and multimodal understanding abilities; (b) deep learning made it possible to learn this task in an 'end-to-end' fashion just from data; (c) obtaining big real-world datasets relied on the crowdsourcing approach; and (d) researchers early on identified various problems particularly related to dataset bias. Most importantly, though, instead of being a practically interesting task in itself, VQA was introduced with the explicit motivation to enable more informative evaluation of multimodal understanding abilities. In this section, I will present a short history of visual question answering, to substantiate the reasons for my choice of focusing on VQA and to put the content of the subsequent chapters into context.

- What object is shining on the animal?
- What objects is the cat sitting behind?
- How many cats?

Figure 3.1: An example image plus associated questions from the VQA Dataset.

Origin. The papers of Malinowski and Fritz (2014a) and Geman et al. (2015) are often seen as the first to, seemingly independently, introduce the VQA task, even though neither of them used the name *"visual question answering"*. A variety of datasets with minor task variations were published in the subsequent year[1], of which the VQA Dataset (Antol et al., 2015) and the name *"visual question answering"* established itself as the standard benchmark. This status has been challenged more recently due to various issues surrounding the dataset and questionable performance of models trained on it.

Motivation. Malinowski and Fritz (2014a) and Malinowski and Fritz (2014b), as well as Geman et al. (2015) and Gao et al. (2015), referred to the VQA task as a *"visual Turing test"*. From a computer vision perspective, visual question answering supersedes pure vision tasks like image classification or object recognition by combining richer vision processing like attribute/relationship/activity recognition and situated language understanding involving referential expressions or commonsense reasoning in a holistic multimodal inference problem. Due to the broad set of abilities which need to be mastered to excel at the task, it was compared to the classic Turing test. Somewhat later, Antol et al. (2015) and Zhu et al. (2016) instead emphasised the advantages of VQA in comparison to image captioning. While both share the focus on multimodal reasoning, a key criticism of image captioning is that valid outputs are relatively unconstrained – there is no single correct caption – which makes it hard to assess performance. In particular, it was found that simple baseline methods are surprisingly effective according to existing evaluation approaches. The strengths of VQA are seen in better capturing the goal of a multimodal benchmark, while being more flexible in particular with respect to the difficulty of instances, and leveraging less problematic automatic evaluation metrics due to its simple output consisting often of just a single word, which 'hides' most of the task complexity. To sum it up, the main motivation for the VQA task has been as a testbed for better evaluation of multimodal understanding abilities.

[1]Time periods (as well as temporal ordering) here and in the following are based on arXiv publication date of the respective papers.

Dataset	Images	Question source	Question types	Performance metric
DAQUAR (Malinowski and Fritz, 2014a)	NYU-Depth2	template-based and human-created	colour, number, object, plus combinations	WUPS
VQA Dataset (Antol et al., 2015)	MS-COCO	crowdsourced	subsequent categorisation: yes/no, number, other	weighted accuracy
COCO-QA (Ren et al., 2015)	MS-COCO	caption-to-question transformations	object, number, colour, location	accuracy
FM-IQA (Gao et al., 2015)	MS-COCO	crowdsourced	subsequent categorisation: 8 types	binary accuracy
Visual Madlibs (Yu et al., 2015)	MS-COCO	template-based	12 types around image and (temporal) context, objects, persons, relationships	accuracy
Visual7W (Zhu et al., 2016)	MS-COCO	crowdsourced	7 w-types: what, where, when, who, why, how, which	accuracy

Table 3.1: Overview of early real-world VQA datasets summarising where images were taken from, how questions were created, what question types are distinguished, and what performance metric is used.

Tasks and datasets. Malinowski and Fritz (2014a) introduced the *"DAtaset for QUestion Answering on Real-world images"* (DAQUAR) dataset, which is based on the NYU-Depth V2 image dataset. Questions and answers are either automatically generated based on templates or created by humans (in-house, not yet crowdsourced), and categorised into a small set of question types (counting, colour, etc). In addition to plain accuracy, they also proposed the WUPS score (Wu-Palmer similarity for sets) as a new performance metric. Geman et al. (2015) did not propose a dataset, but a system based on a query engine proposing an interrogation-style sequence of binary questions for a given image, filtered by a human operator. The VQA Dataset of Antol et al. (2015) is based (mostly) on the MS-COCO image dataset (Lin et al., 2014) and consists entirely of crowdsourced questions (see figure 3.1 for an example). The dataset offers two modes, open-ended answering and multiple-choice from 18 candidate answers, and is accompanied by a detailed analysis of its content, including categorisation into some question types. Ren et al. (2015) introduced the COCO-QA dataset which is obtained via an automatic method of extracting question-answer pairs of certain types (number, location, etc) from given image captions, here based on the MS-COCO captioning dataset. Importantly, they were the first to actively frame VQA as a classification task by reducing answers to a single word – a practice which Antol et al. (2015) used for evaluating their baseline systems as well, and which is subsequently used by most 'users' of the VQA Dataset. Gao et al. (2015) introduced the *"Large-scale Freestyle Multilingual Image Question Answering"* (FM-IQA) dataset, based on MS-COCO, which covers both English and Chinese, and emulates the Turing test setup by letting humans distinguish computer-generated and human answers. The *"Visual Madlibs"* dataset of Yu et al. (2015), again based on MS-COCO, consists of automatically produced fill-in-the-blank templates for multiple-choice questions, which target a range of specific aspects around objects,

persons, their relationship, and more. Finally, the Visual7W dataset of Zhu et al. (2016) is also based on MS-COCO, and focuses on crowdsourced wh-questions (where, who, etc) plus which-questions with visual as opposed to textual answers. See table 3.1 for a summary and comparison of the various VQA datasets.

Models for the VQA Dataset. Antol et al. (2015) and Ren et al. (2015) introduced a wide range of trivial baselines, one leveraging question-type priors, another using nearest neighbours methods, plus image- and question-only models based on a CNN or BoW/LSTM, respectively. Most of the early multimodal models were based on a combination of a pretrained CNN/ResNet module to process the image, an LSTM/GRU/CNN module to process the question word embeddings, sometimes pretrained, and a way of combining both modules before producing the expected output, usually fully-connected layers followed by a classification over possible answers (see Malinowski et al. (2015) and Gao et al. (2015) for answer sequence generation). The processed image embedding may be fed as a first/last 'word' to the RNN (Ren et al., 2015), or concatenated with the word embeddings at every step of the RNN (Malinowski et al., 2015; Gao et al., 2015), but a simple fusion of image and final question embedding via concatenation or pointwise multiplication (Antol et al., 2015; Teney et al., 2018) established itself as the standard CNN-LSTM baseline model for VQA.

In the following years, a variety of extensions and/or modifications to this model were proposed, focusing on the VQA Dataset and encouraged by the annual VQA Challenge (Antol et al., 2015). One class of models improves upon the simplistic fusion operation to allow for more complex interactions between the two modalities, for instance, via convolutions (Ma et al., 2016), multimodal compact bilinear pooling (Fukui et al., 2016), Hadamard low-rank bilinear pooling (Kim et al., 2017) or multimodal Tucker fusion (Ben-Younes et al., 2017). Another common type of improvement is to introduce an attention mechanism, usually over one of the final CNN feature maps and informed by the question embedding, hence another way of fusing both modalities. Examples include basic spatial attention (Xu and Saenko, 2016), multi-step stacked attention (Yang et al., 2016; Kazemi and Elqursh, 2017), attention based on bounding boxes (Shih et al., 2016; Ilievski et al., 2016), or question-image co-attention (Lu et al., 2016). Other models extend the architecture with a memory module (Kumar et al., 2016; Xiong et al., 2016). Another stream of work modifies the baseline architecture more profoundly, by letting the question guide the dynamic assembly of the network from a variety of (reused) compositional modules, referred to as neural module networks (Andreas et al., 2016b; Andreas et al., 2016a).

Problems with the VQA Dataset. Figure 3.2 tracks performance on the VQA Dataset for most of the above referenced models. On the surface at least, the diagram indicates substantial progress towards human-level performance, having started from barely better than the question-only baseline. However, I want to point out a few concerning observations, in particular with

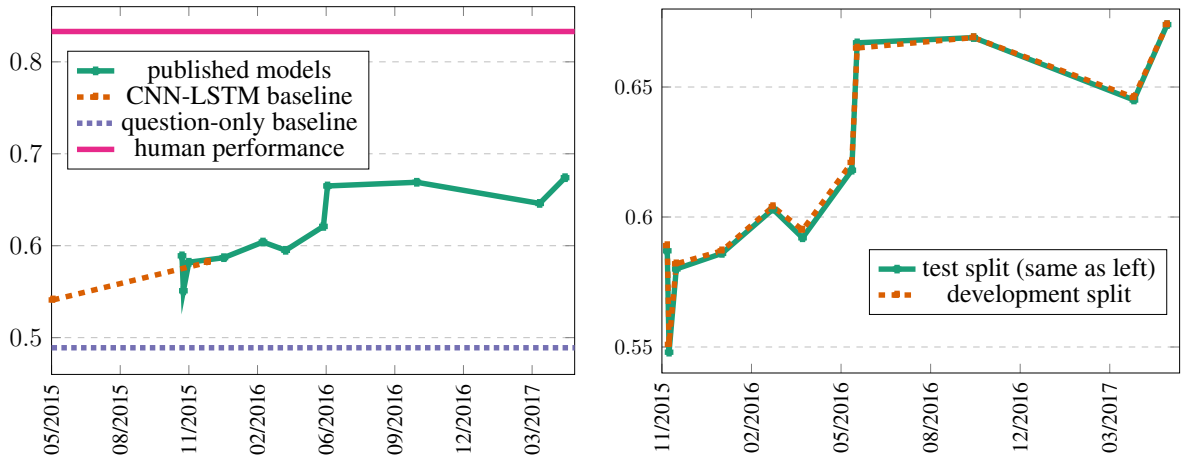


Figure 3.2: Performance over time on the test split of the VQA Dataset, plus a performance comparison on the test vs development split (note different y-axis scale), for a list of models published in the years after its release (*x-axis*: arXiv paper publication date, *y-axis*: accuracy).

regard to the VQA Dataset as an evaluation benchmark. First and foremost, considering that the dataset focuses on multimodal understanding and inference, the supposedly trivial question-only baseline performs surprisingly well, and the human *"ceiling"* performance in turn is suspiciously low. Next, by just tweaking the CNN-LSTM architecture, around 4% improvement could be achieved (Lu et al., 2015), putting this baseline architecture essentially on par with the first wave of VQA models. Most importantly, though, the performance curve gives little hint of the underlying diversity of modelling approaches. Since it is unlikely that the fusion-, attention-, memory- and modularity-focused improvements all have virtually the same effect for the ability to answer visual questions, this indicates the poverty of evaluation detail of the VQA Dataset as monolithic benchmark. In short, it is simply unable to distinguish between these different approaches. Another curious observation is the quasi-consistent slightly worse performance on the development set (0.0–0.3% worse, except in one case), which does not indicate meaningful differences between the diverse set of models with, presumably, different strengths and weaknesses.

Subsequent investigations of the VQA Dataset revealed various data biases and other shortcomings, which may partially explain the aforementioned observations: Zhang et al. (2016) and Goyal et al. (2017) showed strong answer biases when focusing on just the first few words of a question; Agrawal et al. (2016) found that trained models often exhibit only superficial question and image understanding; Jabri et al. (2016) introduced a simple baseline which performs competitively on a slightly modified minimal pairs version of the VQA Dataset setup; Kafle and Kanan (2017b) were concerned with the fact that improvements for simple questions have a much larger impact on performance than for complex ones; Kafle and Kanan (2017a) raised the problem that deciding whether questions are valid is a blind spot of existing evaluation and thus investigated behaviour for absurd questions; Kazemi and Elqursh (2017) indicated

how changing details in the standard CNN-LSTM baseline architecture can be more important than the architecture itself to achieve competitive performance; Chao et al. (2018) analysed the robustness to automatically generated decoys and more generally how to improve the design and collection of interesting datasets for VQA; Mahendru et al. (2017) highlighted the fact that questions come with implicit premises which can help to answer them (like the existence of a *"man"*, a *"racket"* and a *"holding"* relation between them in *"What brand of racket is the man holding?"*); finally, Mudrakarta et al. (2018) pointed out the over-robustness of trained models to semantically meaningful modifications of the question.

To address some of these issues, a range of datasets were introduced which modify or extend the VQA Dataset: a balanced binary yes-no dataset based on parts of the VQA Dataset (Zhang et al., 2016); a language-prior-balanced extension to the VQA Dataset (Goyal et al., 2017) meaning that the phrasing of a question does not favour some answers over other valid candidates; a dataset with richer question types (Kafle and Kanan, 2017a); a compositional train-test split of the VQA Dataset (Agrawal et al., 2017); an extension to a minimal pair setup addressing question premise (Mahendru et al., 2017); a new train-test split with differing answer distributions per question type (Agrawal et al., 2018); and a zero-shot transfer split (Li et al., 2018). While each example addresses an important problem, the 'fixed' dataset is unlikely to avoid other fundamental problems associated with monolithic datasets. Moreover, these solutions are not compatible, as mixing them would make it more difficult to maintain some of the carefully enforced train-test splits. Overall, the series of fixes merely exemplifies the *"vicious cycle"* of dataset creation, as discussed in section 2.5.

CLEVR and other abstract VQA datasets. In this context I started to work on a more principled evaluation approach for deep learning, focusing on the example of visual question answering (Kuhnle and Copestake, 2017). Existing datasets shared some aspects with my proposal, like the use of clipart images (Antol et al., 2015; Zhang et al., 2016) or automatically generated questions (Malinowski and Fritz, 2014a; Ren et al., 2015). Most similar and influential for my approach were the bAbI dataset for reading comprehension (Weston et al., 2015) and the diagnostic SHAPES dataset for VQA (Andreas et al., 2016b), both abstract, fully automatically generated, and specifically designed to evaluate certain capabilities. With the release of the CLEVR dataset (Johnson et al., 2017a) a few months later[2], visual question answering turned into one of the forerunner tasks to explore and adopt such new approaches, within natural language processing at least. CLEVR (Johnson et al., 2017a) is an abstract diagnostic dataset for VQA covering a range of basic multimodal and compositional subtasks, like counting or

[2] To be precise, the CLEVR paper was published on arXiv in December 2016, so after I had submitted my first-year report with the PhD proposal entitled *"Controlled world generation for the evaluation of multi-modal deep learning systems"* in July that year, and had presented a poster on *"Evaluating multi-modal deep learning systems with micro-worlds"* at the Cambridge Language Sciences Annual Symposium in November. Consequently, I developed most of the evaluation approach and the ShapeWorld system simultaneously but independently to CLEVR, and only the later projects in my PhD were influenced by related work based on CLEVR.

comparing numbers of objects and querying or comparing their attributes. Subsequently, Santoro et al. (2017) introduced Sort-of-CLEVR to focus specifically on relational reasoning. The NLVR dataset of Suhr et al. (2017) shares the abstract domain of these datasets, but leverages crowdsourcing to obtain human-produced captions for a VQA-style yes/no caption agreement task. More recently, the COG dataset of Yang et al. (2018) focused on temporal and working-memory-related abilities in a variety of temporal-sequential VQA tasks, and its approach to data generation and evaluation resembles that of CLEVR.

The motivation for these new datasets is generally similar. On the one hand, the issues of the VQA Dataset make it unsuitable for informative evaluation; on the other hand, the advantage of abstract data is seen in: (a) its reduced complexity with respect to visual recognition; (b) the counter-intuitive tendency to encourage more complex instances thanks to its simplistic domain; (c) the focus on specific abilities like relational reasoning or compositional generalisation; (d) and the possibility to control various details of content and representation. While these datasets originated from similar considerations as my proposed evaluation methodology, none of them entirely avoids the fundamental shortcomings of the ML paradigm: the limited flexibility of a single benchmark dataset for in-depth capability-focused evaluation.

Models for CLEVR. The CLEVR dataset has had a big impact on the field, and entailed a series of novel modelling approaches beyond the limited variability of previous VQA architectures. Particularly noteworthy is the fact that the majority of subsequent models were evaluated solely on CLEVR, without validating performance on a real-world benchmark like the VQA Dataset. In natural language processing, the only other recent abstract dataset I am aware of with a similar impact is the bAbI task suite (Weston et al., 2015), whereas relying on only simulations is more common in reinforcement learning, for instance. Hu et al. (2017) and Johnson et al. (2017b) almost simultaneously developed dynamically assembled module networks based on a sequence-to-sequence approach, called N2NMN and PG+EE, respectively. The latter is also the first to surpass *"human performance"* on the dataset – which, however, similar to the VQA Dataset, is suspiciously low[3]. Subsequently, Santoro et al. (2017) presented the RelNet architecture with an explicit architectural prior for relational reasoning (Raposo et al., 2017), and Perez et al. (2018) introduced the FiLM model which leverages feature-wise linear modulations (Dumoulin et al., 2018). Later approaches include the MAC cell (Hudson and Manning, 2018), the DDRprog model (Suarez et al., 2018), the *"transparency-by-design"* approach (Mascharka et al., 2018), and more. Figure 3.3 tracks performance on the CLEVR dataset for these models.

On the one hand, the impact of CLEVR on research for VQA illustrates how improved evaluation data inspires more interesting model development. The numerous models for the

[3]In both cases, human performance corresponds to the accuracy of crowdsourcing workers on a subset of instances, who are neither trained for the task nor encouraged to carefully think about the answer, but instead paid to respond quickly. Note also that the data was originally produced by humans or human-written generation algorithms, so it could be argued that 100% is a more appropriate basis of comparison. In this view, claims of *"(super-)human performance"* could be seen as exaggerated.

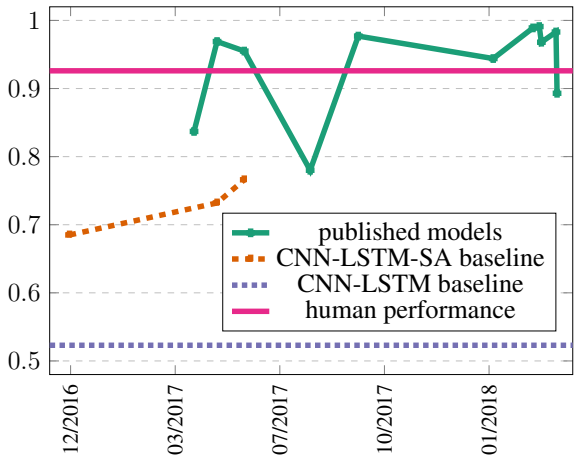


Figure 3.3: Performance over time on the test split of the CLEVR dataset for a list of models published in the years after its release (*x-axis*: arXiv paper publication date, *y-axis*: accuracy).

VQA Dataset followed a few high-level patterns, but were mostly holistic architectures which differed in a range of arbitrary and detailed architectural choices, including baselines that solely focused on tuning such details (Lu et al., 2015; Jabri et al., 2016; Kazemi and Elqursh, 2017). In contrast, most of the models introduced after CLEVR consisted of modular improvements and otherwise generic VQA architectures, without arbitrary hyperparameter choices and extensive tuning. Often, these modules were not fundamentally restricted to the VQA task – for instance, the relational module of Raposo et al. (2017) for relational inference, or the feature-wise modulation of Dumoulin et al. (2018) for fusing (modality) information. However, despite its positive impact, CLEVR is still a static dataset and thus susceptible to similar problems as other monolithic benchmarks, as can be observed more recently: the chase for minimal performance improvements after the $\sim$95% threshold was surpassed, as illustrated in figure 3.3, and the limited ability to distinguish strengths and weaknesses of these different models with close-to-perfect accuracy. Furthermore, CLEVR may be easier than previously thought, considering the competitive performance of the modified early-fusion CNN-LSTM baseline of Malinowski and Doersch (2018). I will analyse some of these state-of-the-art models in chapter 5, and show how a more detailed investigation with targeted data can uncover differences between them, despite their almost equal performance on CLEVR.

Chapter 4

The ShapeWorld system: visually grounded language generation

This chapter presents the ShapeWorld generation system for visually grounded language data, consisting of an image and an accompanying statement about the image[1]. With respect to the overall thesis, ShapeWorld acts as, on the one hand, an illustrative example implementation of a configurable data simulator and, on the other hand, as a diagnostic testbed for the evaluation of visual question answering models in the subsequent experimental chapters.

The full task of visual question answering encompasses a wide range of language-related abilities, some of which may involve resolving ambiguous language, rely on common-sense reasoning, or require background knowledge about certain objects in the image. The evaluation focus of ShapeWorld is reduced to the narrow and comparatively well-defined subset of ***formal-semantics-style*** understanding capabilities, which only relies on literal language interpretation and knowledge-independent reasoning based on the world state as represented by the accompanying image. ShapeWorld aims to resemble such scenarios in the two-dimensional abstract microworld of coloured shapes located on a plane. Importantly, this choice of abstract domain deliberately eliminates most of the lexical complexity of real-world language use and resulting ambiguities, which are otherwise hard to systematically exclude.

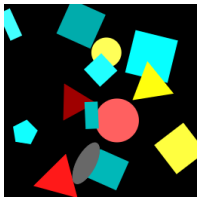
Despite its simplicity, this domain makes it possible to address a range of language understanding abilities, amongst others: objects and their attributes, spatial and other relations between objects, quantified sets of objects and comparisons between them, and logical propositions combining multiple such statements. It is important to keep in mind during the following description of the ShapeWorld system that the aim is not to provide a single dataset, but a configurable generator which makes it possible to produce a variety of fine-grained datasets, each of which provides a complete 'unit-test' benchmark – that is, training and test data – targeting a specific capability. Figure 4.1 illustrates two such datasets for relational and quantification instances but,

[1]The project is open-source and can be found on GitHub: `https://github.com/AlexKuhnle/ShapeWorld`.

A magenta square is to the right of a green shape.

The lowermost green shape is a cross.

A red shape is the same shape as a green shape.



At least half the triangles are red.

More than a third of the shapes are cyan squares.

More than one of the seven cyan shapes is a square.

Figure 4.1: Two example ShapeWorld images with each three accompanying correct or incorrect captions; *left*: relational statements, *right*: quantification statements.

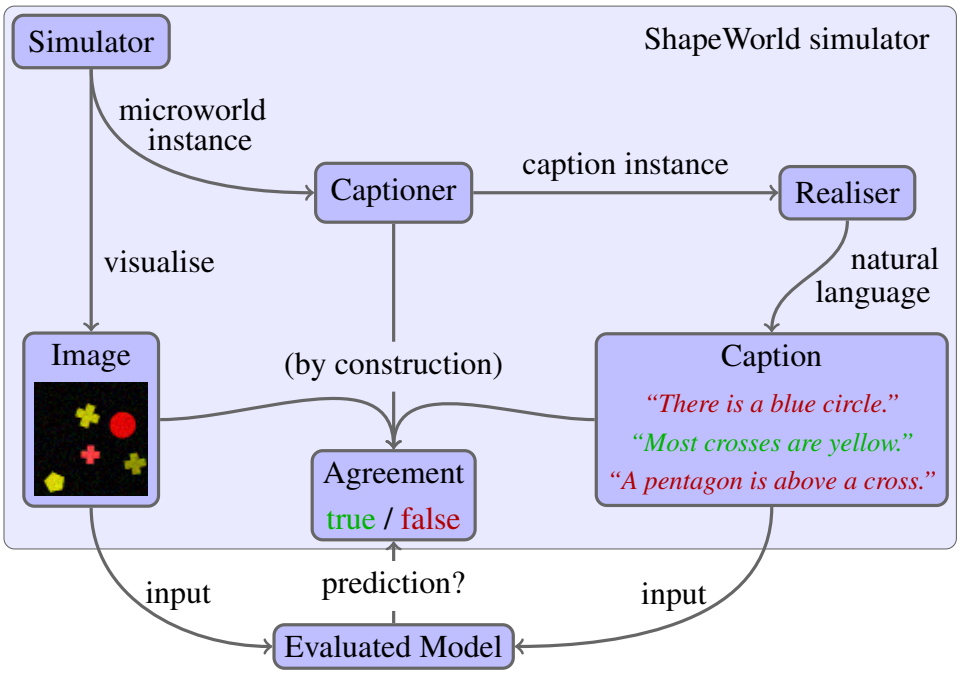


Figure 4.2: High-level overview of the data generation process in the ShapeWorld simulator.

depending on the evaluation focus, other generated datasets may follow either a more targeted or a more generic instance configuration.

Having outlined the evaluation focus (formal-semantics-style language understanding), the microworld (abstract coloured shapes) and the task (binary caption agreement), the following sections will describe the ShapeWorld system and implementation choices related to the unit-testing evaluation focus in more detail. The order of sections reflects the functional order within the generation process: first, the ***simulator module*** randomly samples microworld instances which can be visualised as images (section 4.1), then the ***captioner module*** produces semantic caption representations conditioned on such an instance (section 4.2), and finally the ***realiser module*** transforms caption representations into natural language sentences (section 4.3). Section 4.4 summarises the functioning of the entire architecture. Finally, section 4.5 discusses additional features of a principled generation system for visually grounded language.

Figure 4.2 illustrates the high-level working of the ShapeWorld system. Before going into more detail about the individual modules, I want to emphasise an important aspect of the overall architecture: each of the modules is supposed to be 'generally useful', and their interface is thus chosen to be independent of the rest of the system. Consequently, the captioner produces caption representations solely based on the final output of the simulator, without additionally relying on internal decisions of the simulator sampling process, or informing internal decisions of the caption realiser. The advantage of such modularity is, first, that individual modules can be replaced by other implementations, for instance, to support other languages, as illustrated in section 4.5; second, that the system architecture as a whole can be transferred to other domains, like movie or figure question answering (Tapaswi et al., 2016; Kahou et al., 2018); and third, that parts of the system can be used for other applications, for instance, to produce abstract images for computer vision experiments or caption data for captioning evaluation, as outlined in section 4.5.

4.1 Microworld simulation

A microworld instance is represented as an object-attribute structure which describes all details necessary to visualise the instance as an image.

Objects and attributes. Attributes can be distinguished as either ***primary*** or ***secondary***, where the former is considered semantically meaningful and potentially reflected in a caption, while the latter is only required for visualisation. Primary attributes can further be split into ***absolute*** and ***relative*** attributes, which specify whether their values are semantically meaningful on their own or in relation to another value, respectively. Each attribute is associated with a domain of valid values. Instances are generated via rejection sampling: each object and attribute is iteratively sampled from its domain, and the chosen value is only accepted if it obeys certain global constraints. The following table describes the attributes of a microworld instance:

Attribute	Type	Domain	Typical choice
$*$ size (pixel)	secondary	$x = y > 0$	$x = y = 64$
$*$ background colour[1]	secondary	any colour (see below)	black
$*$ number of objects	primary	$n \geq 0$	$1 \leq n \leq 15$
$*$ objects	primary	list of objects	[see below]
$*$ maximum overlap	secondary	$0.0 \leq o \leq 1.0$	$o = 0.0$ or 0.25
$*$ pixel noise stddev[1]	secondary	$0.0 \leq p \leq 1.0$	$p = 0.0$

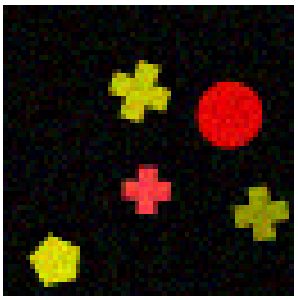
The list of objects is sampled iteratively and then topologically sorted according to overlap, so that, in case of a collision, the object with a relatively larger overlap area is visualised later and hence more in the foreground, to guarantee best-possible visibility. The following table summarises the attributes of an object:

Attribute	Type	Domain	Typical choice
$*$ shape type	absolute	circle, cross, ellipse, pentagon, rectangle, semicircle, square, triangle	all
$*$ shape size (area)	relative	$0.0 < w \cdot h < 1.0$	$0.1 \leq w \cdot h \leq 0.25$
$*$ shape distortion (if asymmetric)	secondary	$d = w/h \geq 1.0$	$2.0 \leq d \leq 3.0$
$*$ colour type	absolute	black, blue, cyan, green, grey, magenta, red, white, yellow	no black and white
$*$ colour shade (shift to black/white)	relative	$-1.0 < s < 1.0$	$-0.4 \leq s \leq 0.4$
$*$ texture[1]	secondary	solid	solid
$*$ centre	relative	$0.0 \leq x, y \leq 1.0$	either uniformly, or in close proximity to another object
$*$ rotation	secondary	$0.0 \leq r < 1.0$	full range
$*$ z-position	relative	determined by global topological sort	automatic
$*$ relative overlap	secondary	computed based on other objects	automatic (≤ 0.25)

An object may be rejected due to invalid collisions with other objects or the boundary of the image. Visually overlapping objects are either disallowed entirely, or a maximum occlusion of 25% of a shape is tolerated. Moreover, collisions between objects of the same colour are only allowed if their shade value differs by at least 0.5, to be able to keep them apart visually.

Simulator configurations. Besides configuring the domain of each attribute, further typical global constraints include: restrictions with respect to the number of objects, to test generalisation to unseen total object numbers, or withholding some shape-colour combinations, to test generalisation to unseen combinations of known attributes. To support such structurally novel data, validation- or test-only conditions can be specified, in which case training data is sampled uniformly from the remaining alternatives. The relative frequency of test values can be specified, so that test instances do not solely consist of a few withheld combinations.

[1]This configuration option has not been used for experiments presented in the thesis since, different from initial expectations, increasing visual variety turned out not to be necessary for sufficiently complex data to obtain interesting experimental results. However, note the discussion on the texture feature in section 5.6.



{ **color**: {**name**: *black*, **shade**: *0.0*}, **noise-stddev**: *0.1*, **size**: *64*, **objects**:
[{ **center**: {**x**: *0.47*, **y**: *0.28*}, **color**: {**name**: *yellow*, **shade**: *-0.24*},
rotation: *0.06*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.10*, **y**: *0.10*}} },
{ **center**: {**x**: *0.49*, **y**: *0.65*}, **color**: {**name**: *red*, **shade**: *0.26*},
rotation: *0.76*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.08*, **y**: *0.08*}} },
{ **center**: {**x**: *0.15*, **y**: *0.91*}, **color**: {**name**: *yellow*, **shade**: *-0.16*},
rotation: *0.27*, **shape**: {**name**: *pentagon*, **extent**: {**x**: *0.09*, **y**: *0.08*}} },
{ **center**: {**x**: *0.80*, **y**: *0.37*}, **color**: {**name**: *red*, **shade**: *-0.12*},
rotation: *0.53*, **shape**: {**name**: *circle*, **extent**: {**x**: *0.12*, **y**: *0.12*}} },
{ **center**: {**x**: *0.92*, **y**: *0.73*}, **color**: {**name**: *yellow*, **shade**: *-0.42*},
rotation: *0.73*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.09*, **y**: *0.09*}} }] }

Figure 4.3: Example output of the ShapeWorld simulator, consisting of a symbolic microworld representation plus its visualisation as image.

In addition, depending on the type of captions which are supposed to accompany the generated image, different sampling strategies are desirable. For later experiments, I mostly use the following additional shape/colour attribute sampling modes:

- Random attributes: both attributes are chosen randomly, hence it is comparatively unlikely to encounter attributes and particularly combinations frequently. This mode is useful for existential and relational statements, where unique attributes are preferred.
- Reinforced attributes: the likelihood of shapes and colours is increased every time they are sampled, which makes it likely to produce certain attributes/combinations more frequently. This mode is useful for number and quantifier statements, where sets of objects with the same attribute(s) are preferred.

These are just examples of typically desirable configuration options, and a wide range of additional parameters is conceivable – see, for instance, the more specialised experiments in section 6.2. The ShapeWorld system shows that the combination of sample domains and rejection sampling based on global constraints supports the efficient implementation of a systematically configurable simulator engine.

Symbolic representation and visualisation. The combined set of primary and secondary attributes fully specifies the appearance of a microworld instance as visual scene, that is, as array of RGB values. The transformation is deterministic with the exception of pixel noise. Internally, a singly-nested dictionary-like attribute-value structure is used to represent the world, and is passed on as argument to the captioner module. The full output of the simulator module includes both the symbolic microworld representation and the corresponding image, as illustrated in figure 4.3.

4.2 Scene captioning

When designing a generator for grounded language like ShapeWorld, a fundamental decision is the order of generation: the ***context-first*** approach produces a language statement conditioned on a given context (here: an image), while the ***language-first*** approach constructs the context with the constraint to reflect the given language statement. Both can be found in the literature, for instance, the CLEVR dataset (Johnson et al., 2017a) takes the former approach whereas the COG dataset (Yang et al., 2018) follows the latter. Besides technical differences in implementation, their relevance lies in avoiding data bias (specifically modality bias, as introduced in section 2.2). Since evaluation data for grounded language understanding is expected to challenge the ability to infer the correct response from language grounded in context, language bias refers to the tendency that task instances can be systematically answered solely based on the language input without considering the context, and context bias vice versa. On the one hand, the context-first approach reduces context bias at the cost of language bias as, for instance, situations representing a specific more complex statement are less likely to randomly emerge than for simple statements. In addition, scenes tend to be less 'constructed' since generation follows the more natural order of language production as reaction to a situation. This, however, can make it hard to guarantee that statements are always fully unambiguous, given arbitrary context configurations. On the other hand, the language-first approach trades less biased language at the cost of more constructed contexts which, by design, unambiguously illustrate the semantics of the given statement.

ShapeWorld follows the context-first approach to generate visually grounded language, for various reasons: first and foremost, since the evaluation focus is on formal-semantics-style tasks, the conditioning order corresponds to the interpretation of language propositions in formal semantics, which is usually formulated with respect to a 'model of the world'. Moreover, although language bias is increased by this choice, the abstract domain of coloured shapes inherently limits the degree of such bias. Rejection sampling and other techniques to further counter specific sources of language bias are discussed in this section. In contrast, the visual context here constitutes a relatively complex space which involves many decisions and consequently, in the case of a language-first approach, would require elaborate construction mechanisms that may introduce a range of intricate and hard-to-pin-down biases. Finally, while ambiguity could also be an interesting language-related evaluation feature, it is not desirable for unit-testing evaluation – unless the investigation specifically targets model behaviour in the context of ambiguity. Once again, the abstract microworld domain and formal-semantics-style statements implicitly reduce the degree and diversity of possible language ambiguities. Remaining ambiguity issues are addressed by basing ShapeWorld semantics on a ternary logical formalism which, besides true and false, explicitly handles ambiguous cases as neither true nor false[2].

[2]The ternary logic is implemented as a more general continuous $[-1.0, 1.0]$ formalism with < 0.0 (usually -1.0) as incorrect, 0.0 as ambiguous and > 0.0 (usually 1.0) as correct. This makes it possible to extend definitions to 'degrees of agreement', however, while the feature is supported, I do not make use of it in experiments.

Section 4.2.1 first introduces the various caption components, which together constitute the domain-specific compositional formal semantics framework for ShapeWorld. The result of the captioning process is a fully-specified proposition, which is independent of a concrete image and instead defines the semantic interpretation of the caption, that is, its agreement with respect to any ShapeWorld image. The captioner module itself is presented in section 4.2.2. It uses a microworld instance as basis for producing such a proposition, and includes a range of bias-reducing as well as performance-optimising mechanisms.

4.2.1 Compositional caption semantics

The caption semantics framework of ShapeWorld is loosely inspired by how language meaning is modelled in formal semantics via a logical formalism. A ***caption component*** can act either as a ***caption*** itself, that is, a fully-specified component which corresponds to a natural language statement or, where applicable, as an argument to another component, compositionally forming a more complex nested structure. The next paragraphs introduce the various caption components which are currently implemented in ShapeWorld. Similar to a microworld, these components are internally represented as dictionary-like attribute-value structures.

Captions and predicates. The interpretation of a caption $[\![c]\!]$ is a function assigning world instances to ternary truth values, $W \longmapsto \{true, false, ambiguous\}$, indicating whether caption c is true, false or ambiguous given a corresponding image. Some caption components – in the following referred to as predicates – are better characterised via two mappings $p.\text{agree}(\cdot)$ and $p.\text{disagree}(\cdot)$, which indicate whether the predicate caption p agrees with an entity or not: $e \longmapsto \{true, false\}$. These two functions alternatively characterise the semantic interpretation of a predicate via:

$$[\![p]\!] = \begin{cases} true & \text{if } \exists\, e \in \mathcal{W} : p.\text{agree}(e) \ [= true] \\ false & \text{if } \forall\, e \in \mathcal{W} : p.\text{disagree}(e) \ [= true] \\ ambiguous & \text{else} \end{cases}$$

Note that an entity may neither clearly agree nor disagree, in which case the truth of the predicate is interpreted as $ambiguous$. Importantly, the boundary to definite truth values is not informed by human perception of ambiguity (a research topic in its own right), but is rather chosen as 'safe margin' which generously excludes potentially controversial configurations. Combinations of a caption and an image with ambiguous agreement are categorically rejected. Consequently, all generated outputs are unambiguously true or false, while there are 'grey zones' of the ShapeWorld instance space which are never sampled.

Attributes. An attribute is a predicate component specified by a type and a value. Available types are: *"shape"*, *"colour"* and *"combination"* which accept as value a corresponding shape/colour value or a pair thereof; and *"shapes"*, *"colours"* and *"combinations"* which accept a set of such values. The following two examples illustrate the general pattern of attribute definitions.

- *"red"* translates to $p = \text{attribute}(\text{type} = \text{colour}, \text{value} = \text{red})$:

$$\begin{aligned} p.\text{agree}(e) &:= e.\text{colour} = \text{red} \\ p.\text{disagree}(e) &:= e.\text{colour} \neq \text{red} \end{aligned}$$

- *"round"* translates to $p = \text{attribute}(\text{type} = \text{shapes}, \text{value} = \{\text{circle}, \text{semicircle}, \text{ellipse}\})$:

$$\begin{aligned} p.\text{agree}(e) &:= e.\text{shape} \in \{\text{circle}, \text{semicircle}, \text{ellipse}\} \\ p.\text{disagree}(e) &:= e.\text{shape} \notin \{\text{circle}, \text{semicircle}, \text{ellipse}\} \end{aligned}$$

Object-types. An object-type is an intersective combination of attributes, itself again forming a predicate component. For instance, *"red square"* combines two attributes to an object-type. Note that an *"object-type"* describes a class of objects (red squares), whereas an *"object"* corresponds to a concrete entity in an image (a concrete red square).

- $p = \text{object-type}(\text{attributes} = A)$:

$$\begin{aligned} p.\text{agree}(e) &:= \forall\, p' \in A : p'.\text{agree}(e) \\ p.\text{disagree}(e) &:= \exists\, p' \in A : p'.\text{disagree}(e) \end{aligned}$$

Relations. A relation is a predicate component specified by a type, a value and a reference object-type. Relations address most properties of an object: x-coordinate (*"to the left/right of X"*), y-coordinate (*"above/below X"*), z-coordinate (*"behind/in front of X"*), shape (*"same/different shape as/from X"*), colour (*"same/different colour as/from X"*), shape size (*"smaller/bigger than X"*), colour shade (*"darker/lighter than X"*), plus two ternary relations addressing relative distances (*"closer to the Y than X"*, *"farther from the Y than X"*) which additionally involve a second unique comparison object-type. Furthermore, both an attribute and an object-type can be trivially turned into a relation (via a form of *"to be X"*, for example, *"is blue"*). The following two examples illustrate the pattern of relation definitions. Note the avoidance of ambiguity by requiring minimal ϵ value differences, by accepting *"left"* only if an object's x-distance is more than its y-distance to the reference, and by accepting *"bigger"* only if both objects have the same shape, since relative size perception can be skewed for certain shape pairs.

- *"to the left of"* translates to $p = \text{relation}(\text{type} = \text{x-rel}, \text{value} = -1, \text{reference} = r)$:

$$\begin{aligned} p.\text{agree}(e) &:= \exists\, e' \in r.\text{agree}(\cdot) : (e'.\text{x} - e.\text{x}) > \max(\epsilon_{\text{distance}}, |e.\text{y} - e'.\text{y}|) \\ p.\text{disagree}(e) &:= \forall\, e' \in \neg r.\text{disagree}(\cdot) : (e'.\text{x} - e.\text{x}) < -\epsilon_{\text{distance}} \end{aligned}$$

- *"bigger"* translates to $p = \text{relation}(\text{type} = \text{area-rel}, \text{value} = 1, \text{reference} = r)$:

$$\begin{aligned} p.\text{agree}(e) &:= \exists\, e' \in r.\text{agree}(\cdot) : (e.\text{area} - e'.\text{area}) > \epsilon_{\text{area}} \ \wedge\ e'.\text{shape} = e.\text{shape} \\ p.\text{disagree}(e) &:= \forall\, e' \in \neg r.\text{disagree}(\cdot) : (e.\text{area} - e'.\text{area}) < -\epsilon_{\text{area}} \end{aligned}$$

Selectors. I refer to a range of *"the... X"* phrases as *"selector"*, like *"the bigger square"* or *"the leftmost circle"*, which 'select' one object from a set of two/multiple objects according to a certain criterion, like size or relative x-location. Each phrase comes in two variations, one based on the positive or comparative form of the adjective, like *"the bigger X"* or *"the left X"* of overall two *"X"*, and another based on the superlative form, like *"the biggest X"* or *"the leftmost X"* of an arbitrary number of *"X"*. Formally, a selector is a predicate component specified by a type, a value and a scope object-type which defines the set of objects from which is selected. Similarly to relations, selectors may address most properties of an object: x-coordinate (*"the left/right X"* and *"the leftmost/rightmost X"*), (*"the upper/lower X"* and *"the uppermost/lowermost X"*), shape size (*"the smaller/bigger X"* and *"the smallest/biggest X"*), colour shade (*"the darker/lighter X"* and *"the darkest/lightest X"*), and two relative distance selectors (*"the X closer/farther to/from the Y"* and *"the X closest/farthest to/from the Y"*) which additionally involve a second unique comparison object-type. The following two examples illustrate the pattern of selector definitions. Similar to the relation example, ambiguity is avoided by requiring minimal differences. Note also that the semantics of the *"the bigger"* example is defined as ambiguous (and consequently discarded in the generation process) unless there are exactly two objects to choose from, as the phrase is not well-defined otherwise.

- *"the bigger"* translates to $p = \text{selector}(\text{type} = \text{area-two}, \text{value} = 1, \text{scope} = s)$:

$$\begin{aligned} p.\text{agree}(e) &:= s.\text{agree}(e) \ \wedge && \text{(part of scope)} \\ &\quad |\{e' : s.\text{agree}(e')\}| = |\{e' : \neg s.\text{disagree}(e')\}| = 2 \ \wedge && \text{(two scope objects)} \\ &\quad \forall^{\geq 1}\, e' \in s.\text{agree}(\cdot) : e' \neq e \ \wedge && \text{(at least one other scope object)} \\ &\quad e'.\text{shape} = e.\text{shape} \ \wedge && \text{(other scope objects have the same shape)} \\ &\quad (e.\text{area} - e'.\text{area}) > \epsilon_{\text{area}} && \text{(other scope objects are smaller)} \end{aligned}$$

$$
\begin{aligned}
p.\text{disagree}(e) \;:=\; & s.\text{disagree}(e) \;\vee && \text{(either not part of scope)} \\
& \Big[|\{e' : s.\text{agree}(e')\}| = |\{e' : \neg s.\text{disagree}(e')\}| = 2 \;\wedge && \text{(or two scope obj.)} \\
& \exists\, e' \in \neg s.\text{disagree}(\cdot) : (e.\text{area} - e'.\text{area}) < -\epsilon_{\text{area}} \Big] && \text{(other objects bigger)}
\end{aligned}
$$

- *"the biggest"* translates to $p = \text{selector}(\text{type} = \text{area-max}, \text{value} = 1, \text{scope} = s)$:

$$
\begin{aligned}
p.\text{agree}(e) \;:=\; & s.\text{agree}(e) \;\wedge && \text{(part of scope)} \\
& |\{e' : s.\text{agree}(e')\}| \geq 2 \;\wedge && \text{(at least two scope objects)} \\
& \forall^{\geq 1}\, e' \in s.\text{agree}(\cdot) : e' \neq e \;\wedge && \text{(at least one other scope object)} \\
& e'.\text{shape} = e.\text{shape} \;\wedge && \text{(other scope objects have the same shape)} \\
& (e.\text{area} - e'.\text{area}) > \epsilon_{\text{area}} && \text{(other scope objects are smaller)} \\
p.\text{disagree}(e) \;:=\; & s.\text{disagree}(e) \;\vee && \text{(either not part of scope)} \\
& \Big[|\{e' : s.\text{agree}(e')\}| \geq 2 \;\wedge && \text{(or at least two scope objects)} \\
& \exists\, e' \in \neg s.\text{disagree}(\cdot) : (e.\text{area} - e'.\text{area}) < -\epsilon_{\text{area}} \Big] && \text{(other objects bigger)}
\end{aligned}
$$

Existentials. An existential is a combination of an object-type or selector acting as subject, and a relation acting as verb.

- $c = \text{existential}(\text{subject} = s, \text{verb} = v)$:

$$
[\![c]\!](W) \;:=\; \begin{cases} \mathit{true} & \text{if } \exists\, e \in W : s.\text{agree}(e) \wedge v.\text{agree}(e) \\ \mathit{false} & \text{if } \forall\, e \in W : s.\text{disagree}(e) \vee v.\text{disagree}(e) \\ \mathit{ambiguous} & \text{else} \end{cases}
$$

Quantifiers. A quantifier is a caption component specified by a type (*"count"* or *"ratio"*), a comparator (*"equal"*, *"not equal"*, *"less than"*, *"at most"*, *"more than"*, *"at least"*), a quantity, plus an object-type acting as subject and a relation as verb. The type defines whether object numbers are quantified, like *"three"*, or fractions between set cardinalities, like *"half"*[3]. The quantity specifies the reference number/fraction, which combined with the comparator yields the associated truth value. Currently supported quantities are the numbers 0 to 5, and 0.0 (*"no"*), 0.25 (*"a quarter of"*), 0.33 (*"a third of"*), 0.5 (*"half"*), 0.66 (*"two thirds of"*), 0.75 (*"three quarters of"*), and 1.0 (*"all"*). Trivial and nonsensical combinations are excluded, for instance, *"less than/at most zero/no"* or *"more than/at least all"*. The following two examples illustrate the

[3]Note that the current definition avoids non-trivial nested quantification, which simplifies some of the definitions, but an extension to 'sets-of-sets' semantics required to handle arbitrarily nested quantifiers is straightforward.

pattern of quantifier definitions. Once again, note the avoidance of ambiguity by conservatively using the smaller set of agreeing versus the larger set of not-disagreeing objects to decide for agreement, and vice versa for disagreement.

- *"at most three"* translates to $c = \text{quantifier}(\text{type} = \text{count}, \text{comparator} = \text{at most}, \text{quantity} = 3, \text{subject} = s, \text{verb} = v)$:

$$[\![c]\!](W) \quad := \quad \begin{cases} \mathit{true} & \text{if } |\{e \in W : \neg s.\text{disagree}(e) \wedge \neg v.\text{disagree}(e)\}| \leq 3 \\ \mathit{false} & \text{if } |\{e \in W : s.\text{agree}(e) \wedge v.\text{agree}(e)\}| > 3 \\ \mathit{ambiguous} & \text{else} \end{cases}$$

- *"most"* / *"more than half"* translates to $c = \text{quantifier}(\text{type} = \text{ratio}, \text{comparator} = \text{more than}, \text{quantity} = 0.5, \text{subject} = s, \text{verb} = v)$:

$$[\![c]\!](W) \quad := \quad \begin{cases} \mathit{true} & \text{if } \frac{|\{e \in W : s.\text{agree}(e) \wedge v.\text{agree}(e)\}|}{|\{e \in W : \neg s.\text{disagree}(e)\}|} > 0.5 \\ \mathit{false} & \text{if } \frac{|\{e \in W : \neg s.\text{disagree}(e) \wedge \neg v.\text{disagree}(e)\}|}{|\{e \in W : s.\text{agree}(e)\}|} \leq 0.5 \\ \mathit{ambiguous} & \text{else} \end{cases}$$

Propositions. A proposition is a caption component specified by a type and a set of clause captions. It combines the truth values of these clauses according to common logical operators given by the type: *"conjunction"*, *"disjunction"*, *"implication"* (requiring exactly two clauses), and *"equivalence"*. The following two examples illustrate the pattern of proposition definitions. Note that these definitions use the fact that the ternary logic formalism is implemented as continuous values between -1.0 and 1.0.

- *"and"* translates to $c = \text{proposition}(\text{type} = \text{conjunction}, \text{clauses} = C)$:

$$[\![c]\!](W) \quad := \quad \min_{p \in C} [\![p]\!](W)$$

- *"if and only if"* translates to $c = \text{proposition}(\text{type} = \text{equivalence}, \text{clauses} = C)$:

$$[\![c]\!](W) \quad := \quad \max\left(\min_{p \in C} [\![p]\!](W),\ \min_{p \in C} -[\![p]\!](W)\right)$$

4.2.2 Caption sampling mechanism

The currently supported set of general-purpose captioners makes use of the compositional caption system to generate fully-specified output captions. A captioner corresponds to one or multiple parametrised caption component 'templates', for instance, the component-equivalent of *"A [(colour) (shape)] is [relation] a [(colour) (shape)]."*. Captioning in ShapeWorld is a two-step

process: first, the captioner modules are initialised by randomly sampling parts of the component values independently, and subsequently the specification of the output caption is completed based on a concrete microworld instance. If the generated caption is supposed to be wrong with respect to the image, an additional step invalidates the correct statement, usually by changing a single detail of the correct caption to guarantee a plausible and minimally incorrect statement.

The microworld-independent initialisation prevents the distributions of component values from being context-biased, as otherwise patterns originating from the image generation would result in corresponding caption biases. For instance, one more frequently encounters smaller numbers of more specific object descriptions, like *"one red square"*, and larger numbers of less specific ones, like *"four shapes"*, and an image-conditioned choice of values would reflect this pattern. In fact, most of the caption values are potentially affected by this (quantifiers, relations, logical connectives, etc), hence the only values sampled conditioned on the image are the values of shape and colour attributes. Additionally, the choice of whether to produce an incorrect caption, and if so, what caption component to invalidate, is decided as part of the initialisation, not based on the concrete image.

In the course of the context-independent captioner initialisation, the logical predication resulting from the sampled values is analysed and implications on the resulting caption semantics recorded. This is done to be able to control whether semantic redundancies, tautologies or contradictions (see below) are accepted when assembling the caption object, ultimately addressing language bias like trivial statements that are always true/false. For instance, if an object property is referred to in one component, then this may duplicate information already given in another component. The following list illustrates the three phenomena:

- Logical redundancy: *"A square is a red square."* (shape redundancy in object, but additional colour information)
- Logical tautology: *"A square is a square."* (shape redundancy in object, no additional information)
- Logical contradiction: *"A square is a red (*square*) circle."* (shape redundancy in object, but invalidated, hence resulting in a contradictory incorrect statement)

Relations like *"is the same shape as"* or *"is a different colour from"* are also taken into account here. While the analysis could in principle be extended to a full SAT-solver-based logical analysis, considering all properties and implications including spatial and number relations, the current system only implements a simplified mechanism focusing on repetitive shape/colour specifications, as other problems are unlikely to be produced given the relatively basic caption patterns used so far.

Similarly, the context-dependent 'pragmatic predication' is analysed during the process of the second stage of world-conditioned caption completion. This addresses semantic redundancies

originating from pragmatic considerations, which lead to trivialities regarding visual bias where, given a scene, a statement can be identified as true/false while ignoring parts of it. For instance, *"red"* in *"a red square..."* is pragmatically redundant if there are only red squares, since no colour distinction is required then. Pragmatic predications keep track of the set of agreeing objects of sub-expressions, and prevent caption elements which do not effectively reduce this set and thus not add relevant additional information.

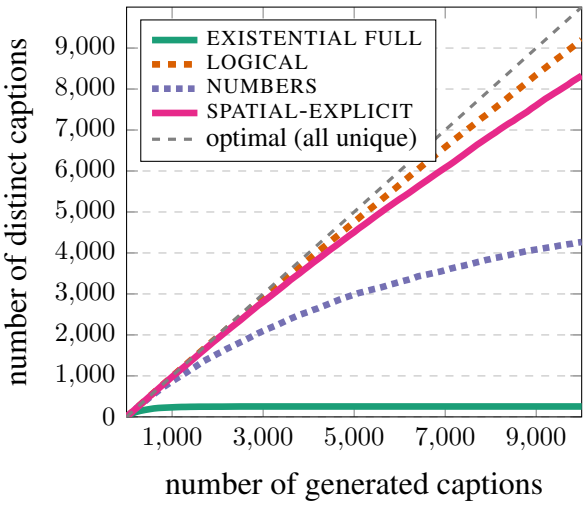
By default, both logical and pragmatic redundancies are accepted, while logical tautologies and contradictions are rejected. This eliminates the most serious cases of trivial language-only instances, while still supporting varied patterns of object references.

4.2.3 Key design choices

The following list summarises the crucial captioner design choices to reduce bias and ambiguity, thus improving data quality and, ultimately, the quality of evaluation:

- Context-first approach to generation, to resemble formal semantics setup and avoid hard-to-pin-down visual biases.
- Compositional caption representation, for a systematic sampling process and a combinatorial diversity of captions.
- Explicit modelling and subsequent avoidance of semantically ambiguous instances, by employing a ternary logic and constraining semantics via safe margins.
- Two-stage sampling approach, with unbiased initialisation of values which may otherwise lead to language bias, and context-conditioned completion of others.
- Logical and pragmatical predication analysis, to avoid semantic trivialities.
- Incorrect captions as invalidated correct captions, to guarantee the plausibility of incorrect statements and the minimal-pair-style difficulty of distinguishing them from correct ones.

To illustrate the rich variety of captions that the outlined approach produces for different statement types, the diagram to the right indicates the number of distinct captions in a set of generated captions for different datasets (details not important here, but see section 5.3 for definition of the datasets).



4.3 Caption realisation

While the visualisation of an abstract microworld model is straightforward, the process of turning a caption model representation into natural language is less obvious. Central to my approach in ShapeWorld is the use of the semantic representation framework of Minimal Recursion Semantics (MRS) (Copestake et al., 2005) and its dependency graph variant DMRS (Copestake, 2009), in combination with technology made available by the DELPH-IN (Deep Linguistic Processing with HPSG) consortium. First, the English Resource Grammar (ERG) (Flickinger, 2000; Flickinger et al., 2014) is a bi-directional, broad-coverage and high-precision implementation of MRS for the English language. It is linguistically precise, meaning that sentences only parse if they are valid according to its hand-built rules, and general-purpose, with a verified coverage of around 80% for Wikipedia, and over 90% for corpora with shorter sentences and more limited vocabulary (for a detailed discussion see (Flickinger, 2011)). Second, the Answer Constraint Engine (ACE) (Packard, 2018) is a parser-generator for MRS-based grammars, allowing to both efficiently parse sentences into MRS as well as generate natural language from MRS representations.

Section 4.3.1 presents (D)MRS in more detail and motivates its choice as a suitable representation for ShapeWorld and similar data simulator frameworks. Section 4.3.2 describes how ShapeWorld caption components are mapped to DMRS graph snippets and composed to a full valid DMRS graph, which includes my contributions of DMRS matching, rewriting, and a simple DMRS description language to the pydmrs framework (Copestake et al., 2016).

4.3.1 Dependency Minimal Recursion Semantics

MRS (Copestake et al., 2005) is a logical formalism which is particularly suited to model the formal semantics of natural language. It follows the typical approach of distinguishing between ***instance*** and ***event variables***, whose contents and relations are specified via ***predications***. Instance and event variables furthermore have language-specific attributes like *"number"* or *"tense"*. For instance, the meaning of exampleA square is red. may be modelled as follows:

$$\exists\, e[\text{tense}\colon \text{present}, \dots\,]\ \exists\, x[\text{number}\colon \text{singular}, \dots\,] : \text{red}(e) \wedge \text{square}(x) \wedge \text{ARG1}(e, x).$$

Different from many other logical semantic formalisms, MRS introduces another type of variable, ***scope handles***, and represents the semantics not as a linear logical formula, but as a set of predication objects. Each predication is associated with a scope handle and introduces either an instance or event variable plus, where applicable, specifies its relation arguments (mostly for event variables). Arguments can either directly refer to variables or handles ***Quantifiers*** are modelled as a special type of predication which does not introduce its own variable, but via underspecified scoping is associated with the instance variable it quantifies. The example above may consequently be modelled as follows:

$$\big\{[h_1 : \text{red}(e_1[\ldots]),\ \text{ARG1}(e_1, x_1)],\ [h_2 : \text{square}(x_1[\ldots])], [h_3 : \text{a}(x_1), \text{RSTR}(h_4 \cong h_2)]\big\}$$

Note that the quantifier scope is underspecified as h_4 instead directly pointing to h_2. If an argument introduces a new handle, additional ***handle constraints***, in this case $h_4 \cong h_2$, specify scoping restrictions which allow to recover the correct (and only the correct) logical interpretation(s) of the statement's formal semantics. Handle constraints become important, for instance, if a sentence contains both existential and universal quantifiers, like *"All squares are above a circle."*. Here, each handle constraint would associate the quantifier with its corresponding noun predicate, while the order in which these handle constraints are resolved is underspecified:

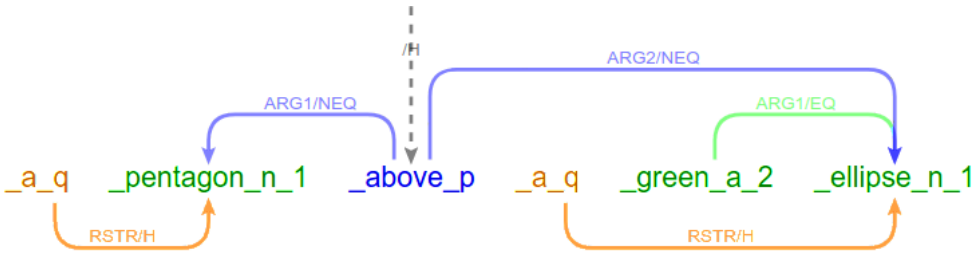
$$\forall\, x_1 : \text{square}(x_1) \wedge \big[\exists\, x_2 : \text{circle}(x_2) \wedge \ldots\big] \quad \text{vs} \quad \exists\, x_1 : \text{circle}(x_1) \wedge \big[\forall\, x_2 : \text{square}(x_2) \wedge \ldots\big]$$

Overall, an MRS representation consists of a set of predications and handle constraints, plus top handle and index variable specifying outermost scope and predication (see figure 4.4). Importantly, the MRS formalism is inherently compositional, and defines the rules for constructing an MRS structure as an iterative process of adding predicates and incorporating argument as well as scoping relations.

DMRS (Copestake, 2009) uses the fact that MRS structures with minor additional constraints can equivalently be represented as a directed acyclic graphs, where each node corresponds to a variable and its predicate, while argument relations and handle constraints are expressed as labelled edges between nodes. See figure 4.4 for a full example of an MRS structure, corresponding DMRS graph, plus associated English sentence. The formalism is very similar to another MRS variant, Elementary Dependency Structures representation (Oepen and Lønning, 2006), however, while the latter strives for simplicity and comparability with syntactic dependency formalisms at the cost of structural information loss, meaning it is not possible to recover the original MRS structure from an EDS graph, the former is fully interconvertible with MRS and thus better suited for a setup which requires going from a high-level semantic representation like DMRS graphs via its lower-level MRS equivalent to natural language output.

As becomes clear from the description of MRS, the semantic caption representation of ShapeWorld is essentially a simplified version of this formalism[4]. By mapping caption components to DMRS representations as described in section 4.3.2, the resulting semantic graph acts like a (partial) language-specific annotation of the underlying microworld, with nodes/predicates

[4]I want to acknowledge Woodley Packard's demo project for a tutorial on *"English Resource Semantics"* at LREC and NAACL 2016 here since, despite being unrelated to evaluation or deep learning, it was one of the inspirations for my approach. Similarly to ShapeWorld, this demo checked the truth of statements about an abstract image by parsing the language input via MRS to SQL queries for the underlying symbolic world model representation.




```
[ LTOP: h0 INDEX: e2 RELS: <
    [ _a_q LBL: h4 ARG0:  x3 RSTR: h5 BODY: h6 ]
    [ _pentagon_n_1 LBL: h7 ARG0:  x3 [...]  ]
    [ _above_p LBL: h1 ARG0:  e2 [...]  ARG1:  x3 ARG2:  x8 ]
    [ _a_q LBL: h9 ARG0:  x8 RSTR: h10 BODY: h11 ]
    [ _green_a_2 LBL: h12 ARG0:  e13 [...]  ARG1:  x8 ]
    [ _ellipse_n_1 LBL: h12 ARG0:  x8 [...]  ]
  > HCONS: < h0 ≅ h1, h5 ≅ h7 h10 ≅ h12 > ]
```

"A pentagon is above a green ellipse."

Figure 4.4: MRS structure and corresponding DMRS graph for a simple sentence.

corresponding either to objects and their attributes (*"square"*, *"red"*), or to relations between objects (*"above"*, *"bigger"*), or to propositions about the entirety of objects in the scene (*"at most three"*, *"X and Y"*). For this correspondence and its efficient implementation, the compositionality of the DMRS formalism is crucial: it makes it possible to define the mapping to DMRS graph snippets on the component level, and construct the full graph iteratively by composing the snippets.

Fundamentally, my approach to natural language generation and the role of a compositional formalism for semantic representations within it follows the framework of Bender et al. (2015): on the one hand, DMRS acts as the general-purpose natural language interpretation layer which abstracts away language-specific and context-independent grammatical peculiarities (and only those!), and offers instead a semantic interface which facilitates further processing; the ShapeWorld caption components, on the other hand, constitute an additional domain-specific layer which defines the actual logical semantics of the microworld while ignoring aspects of the general-purpose representation which are irrelevant to the application. I emphasise, in unison with Bender et al. (2015), the comprehensiveness, consistency and scalability of such an approach to language generation – particularly in an abstract domain – as opposed to the frequently seen ad hoc approach based on language templates, for instance, in the case if the bAbI tasks (Weston et al., 2015) or CLEVR (Johnson et al., 2017a). New shape types simply require the addition of an appropriate DMRS graph snippet, while new caption patterns only require the definition of a corresponding component, without needing to adapt or modify existing components. Even completely changing the abstract domain from coloured shapes to, for instance, the clipart domain of Zitnick et al. (2016) does not involve more than introducing novel attributes and

relations, while keeping the overall infrastructure and already existing caption component types. In contrast, template-based approaches are prone to struggle even with minor language aspects as simple as the correct usage of *"a"* versus *"an"* in English.

4.3.2 Mapping, composition and paraphrasing

Compositionality greatly simplifies the generation process by dividing it into two steps: first, the individual components are mapped to corresponding DMRS graph snippets, and subsequently these snippets are recursively combined. In addition, a paraphrasing step post-processes the resulting DMRS graph, before it is passed on to ACE/ERG to be turned into natural language. All of these steps make use of the DMRS graph description language I developed, ***GraphLang*** in the following, which will consequently be presented first.

GraphLang. The GraphLang formalism describes DMRS graphs in a serialised form, as a list of edge sequences plus connected nodes such that every edge of the graph appears exactly once. Nodes are represented by their predicate, e.g. `_square_n_1`, and associated variable with attributes (optionally abbreviated) in square brackets, e.g. `x[3s_+_]`[5], so the node for *"square"* is denoted by `_square_n_1 x[3s_+_]`. The various edge relation types are represented by respective shortforms, for instance, `=1=>` for ARG1/EQ (e.g. a relation between adjective and noun), `-2h->` for ARG2/H (e.g. a relation between verb and gerund object), or simply `-->` for RSTR/H, which is the special relation connecting quantifier and quantified instance node. Finally, graph-level index/top nodes are indicated by a leading `*` and `**`, respectively (often the ERG can infer one from the other). The sentence *"A square is red."* as DMRS graph is thus written as follows:

```
_a_q --> _square_n_1 x[3s_+_] <-1- *_red_a_1 e[ppi--]
```

A node may need to appear multiple times in the linearised representation, for instance, if it is part of more than two edges. Instead of duplicating the node definition in such cases, it is referred to either by a leading colon plus predicate name (if unique), e.g. `:_square_n_1`, or by introducing a reference label, e.g. `subj:_square_n_1`, and referring to it via leading colon plus label instead. The sentence *"A big square is red."* (*"square"* is part of three edges) can hence be written as follows:

```
_a_q --> subj:_square_n_1 x[3s_+_] <-1- *_red_a_1 e[ppi--];
                :subj <=1= _big_a_1 e[pui--]
```

[5] Description of variable attributes here and in the following (for a full specification, see Kuhnle (2016)): `x[3s_+_]`: 3rd person, singular, individuated; `e[pui--]`: proposition, untensed, indicative; `e[ppi--]`: proposition, present, indicative.

GraphLang supports a variety of advanced features to enable its use for subgraph matching, rewriting and querying. The full list of features can be found in the specification (Kuhnle, 2016) and additional example applications in the pydmrs paper (Copestake et al., 2016). Relevant concepts for the remainder of the section are underspecification and anchor nodes. First, the question mark symbol serves as a universal underspecified marker and can be used in a variety of places: for predicate slots, e.g. `_?_n_1` (any noun predicate); for variables, e.g. `x?` (instance variable with any person/number/etc); individual variable attributes, e.g. `x[3??+?]` (any 3rd person and individuated instance variable); or edges, e.g. `-?->` (any relation type). The special values `pred` and `node` signal a fully underspecified predicate or node (predicate plus variable), respectively. Second, anchor nodes are defined via square-bracketed labels, e.g. `[subj]:_square_n_1 x?`. They can be explicitly referred to by, for instance, rewriting algorithms, and are useful in combination with underspecification, as illustrated in the following.

Mapping. A lookup table maps caption components based on their values to DMRS graph snippets, specified in GraphLang format. The following illustrates this mapping with one example per caption type taken from section 4.2:

- Attribute *"red"*:

```
[attr]:_red_a_1 e? =1=> [type]:node <-- [quant]:default_q
```

- Object-type *"shape"*:

```
[type]:_shape_n_1 x?[pers=3] <-- [quant]:default_q
```

- Relation *"to the left of"*:

```
_left_n_of x?[num=sg] -1-> [ref]:node <-- [quant]:_a_q;
[rel]:_to_p e? -2-> :_left_n_of <-- _the_q
```

- Selector *"the bigger"*:

```
[sel]:_big_a_1 e? =1=> [type]:pred x?[num=s] <-- [quant]:_the_q;
more_comp e? =1=> :sel
```

- Existential:

```
[quant]:_a_q --> [rstr]:pred x?[num=sg] <-1- [body]:node
```

- Quantifier *"at most three"*:

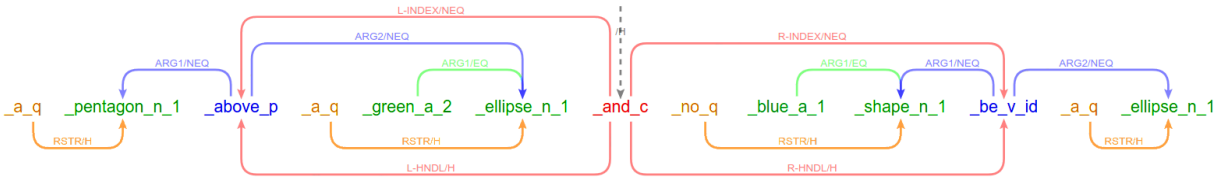
```
[quant]:udef_q --> [rstr]:pred x?[num=pl] <-1- [body]:node;
:rstr <=1= card(3) e? <=1= _at+most_x_deg e?
```

- Proposition *"and"*:

```
[arg1]:node <=1= *[head]:_and_c e[ppi--] =2=> [arg2]:node
```

"A pentagon is above a green ellipse, and no blue shape is an ellipse."

⇑ ACE + ERG realisation ⇑



⇑ Component DMRS mapping ⇑

$\exists a$	a.shape=pg	a.y>b.y	$\exists b$	b.color=gr	b.shape=el	$\wedge$	$\neg\exists c$	c.color=bl	true	$c{=}d$	$\exists d$	d.shape=el
$\exists a:$ a.shape=pg		a.y>b.y	$\exists b:$ b.color=gr $\wedge$ b.shape=el			$\wedge$	$\neg\exists c:$ c.color=bl			$c{=}d$	$\exists d:$ d.shape=el	
$\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]						$\wedge$	$\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c{=}d$]					
($\exists a:$ a.shape=pg $\wedge$ [$\exists b:$ b.color=gr $\wedge$ b.shape=el $\wedge$ a.y>b.y]) $\wedge$ ($\neg\exists c:$ c.color=bl $\wedge$ [$\exists d:$ d.shape=el $\wedge$ $c{=}d$])												

Figure 4.5: A sentence, its associated DMRS graph with coloured components, and a simplified version of the corresponding ShapeWorld semantics, illustrating its compositionality.

Composition. The DMRS snippets obtained this way are iteratively composed. This merging process is guided by the anchor nodes, which act as the glue points for combining child with parent DMRS graphs. Partial underspecification of anchor nodes is resolved by adopting the more specific value, while other non-anchor nodes are simply copied. The unification of anchor predicates takes into account customised predicate hierarchies. For instance, `default_q` subsumes all quantifier predicates, and `_shape_n_1` may act as hypernym for concrete shape predicates like `_square_n_1`. Figure 4.5 illustrates the correspondence between caption component semantics and DMRS graph snippets as well as the compositional structure of the caption.

Matching and rewriting. In addition to composing DMRS graphs, anchor nodes also serve as reference points for DMRS subgraph rewriting. This is a two-step process, consisting of the identification of a subgraph and subsequent replacement by another, to obtain a modified DMRS graph. First, the subgraph S to be replaced has to be located in the DMRS graph G. This is achieved by subgraph matching, where nodes in S are associated with corresponding nodes in G such that all edges match as well. Matching also supports underspecification, as otherwise generic rewriting rules – like *"a [colour] [shape]"* to *"a [shape] which is [colour]"* – would require to enumerate all possible concrete instantiations. Finally, the identified subgraph S within G is transplanted and replaced by another subgraph S' based on the correspondence of their anchor nodes which act as glue points, similar to their role during composition above.

Paraphrasing as subgraph rewriting constitutes the final step in the caption realisation pipeline of the ShapeWorld system, before the DMRS graph is turned into natural language. On the one hand, paraphrasing may be necessary to 'fix' certain technical inconsistencies between ShapeWorld and DMRS semantics, due to grammar-incompatible simplifications in the Shape-

World semantics. For instance, a sentence like *"A square is red."* is internally produced as *"A square is a red shape."*, due to the compositional caption system which pairs adjectives like *"red"* with the semantically empty *"shape"*. However, in English it is more common to collapse a sentence like *"A square is a red shape."* to *"A square is red."*, which can be adjusted by suitable paraphrasing rules. On the other hand, such rules can increase the linguistic variety of vocabulary and constructions by specifying semantically equivalent formulations as sub-graph alternatives, ranging from word-level synonyms like referring to *"red object"* instead of *"red shape"*, to phrase-level synonyms such as paraphrasing *"most squares"* as *"more than half of the squares"*, to clause-level equivalences like *"a shape is red"* and *"there is a red shape"*. Note that the current version of ShapeWorld does not implement instances of this second type of paraphrasing rules since, different from initial expectations, increasing linguistic variety turned out not to be necessary for sufficiently complex data to obtain interesting experimental results.

4.4 System summary: step-by-step overview of ShapeWorld data generation

The purpose of this section is to succinctly summarise the data generation process and point out the various mechanisms to avoid biased and ambiguous data. While figure 4.2 at the beginning of this chapter presented a superficial outline, the following listing incorporates much of the details of the last sections in a more pseudocode-style enumeration of the generation steps and subroutines, with an example illustrating each step.

1. Simulate and visualise microworld instance (section 4.1):

   (a) Initialise global attributes: number of objects, etc

   { **color**: {**name**: *black*, **shade**: *0.0*}, **noise-stddev**: *0.1*, **size**: *64*, **num-objects**: *5* }

   (b) Iteratively sample objects, choosing attributes randomly from their domains, possibly with changing distributions and/or restrictions for training and test data

   Repeat until **num-objects** = 5:
   center ∼ uniform[0.0, 1.0], **rotation** ∼ uniform[0.0, 1.0]
   shape ∼ uniform[square, rectangle, triangle, circle,...], **extent** ∼ uniform[0.1, 0.25]
   color ∼ uniform[red, green, blue, yellow,...], **shade** ∼ uniform[-0.4, 0.4],
   such that (**shape**, **color**) ∉ {(square, red), (triangle, green),...} and no invalid collisions

   (c) Reject collisions with any/high degree of overlap

   (d) Visualise microworld model as image

   (e) Return symbolic representation

{ **color**: {**name**: *black*, **shade**: *0.0*}, **noise-stddev**: *0.1*, **size**: *64*, **objects**:
[{ **center**: {**x**: *0.47*, **y**: *0.28*}, **rotation**: *0.06*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.10*, **y**: *0.10*}}, **color**: {**name**: *yellow*, **shade**: *-0.24*} },
{ **center**: {**x**: *0.49*, **y**: *0.65*}, **rotation**: *0.76*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.08*, **y**: *0.08*}}, **color**: {**name**: *red*, **shade**: *0.26*} },
{ **center**: {**x**: *0.15*, **y**: *0.91*}, **rotation**: *0.27*, **shape**: {**name**: *pentagon*, **extent**: {**x**: *0.09*, **y**: *0.08*}}, **color**: {**name**: *yellow*, **shade**: *-0.16*} },
{ **center**: {**x**: *0.80*, **y**: *0.37*}, **rotation**: *0.53*, **shape**: {**name**: *circle*, **extent**: {**x**: *0.12*, **y**: *0.12*}}, **color**: {**name**: *red*, **shade**: *-0.12*} },
{ **center**: {**x**: *0.92*, **y**: *0.73*}, **rotation**: *0.73*, **shape**: {**name**: *cross*, **extent**: {**x**: *0.09*, **y**: *0.09*}}, **color**: {**name**: *yellow*, **shade**: *-0.42*} }] }

2. Produce (correct) caption for instance (section 4.2):

   (a) Independently initialise parts of captioner values: caption structure, most caption values, incorrect mode (invalidation of captions in step 3)

   { **component**: *existential*, **incorrect-mode**: *body*,
   restrictor: { **component**: *object-type*, **attributes**: [{**shape**: *???*}], **incorrect-mode**: *none* },
   body: { **component**: *type-relation*,
   type: { **component**: *object-type*, **attributes**: [{**shape**: *???*}, {**color**: *???*}], **incorrect-mode**: *color* }
   }}

   (b) Analyse logical structure: redundancy (allowed by default), tautology, contradiction

   Redundancy valid: **restrictor**: **attributes**: *shape* and **body**: **attributes**: *shape*
   Contradition invalid: **restrictor**: **attributes**: *shape* and **body**: **incorrect-mode**: (!) *shape*

   (c) Sample remaining captioner values conditioned on microworld model: mainly concrete values for shape and colour predicates

   restrictor: **attributes**: [{**shape**: *circle*}]
   body: **type**: **attributes**: [{**shape**: *circle*}, {**color**: *red*}]

   (d) Check for pragmatical redundancy (allowed by default)

   (e) Guarantee well-defined agreement via use of ternary logic and exclusion of ambiguous cases

   (f) Return correct caption model

   { **component**: *existential*, **incorrect-mode**: *body*,
   restrictor: { **component**: *object-type*, **attributes**: [{**shape**: *circle*}], **incorrect-mode**: *none* },
   body: { **component**: *type-relation*,
   type: { **component**: *object-type*, **attributes**: [{**shape**: *circle*}, {**color**: *red*}], **incorrect-mode**: *color* }
   }}

3. If caption to be generated is supposed to be incorrect:

   (a) Invalidate correct caption model, usually by modifying a single detail to keep it minimally different and maximally plausible

   { **component**: *existential*, **incorrect-mode**: *body*,
   restrictor: { **component**: *object-type*, **attributes**: [{**shape**: *circle*}], **incorrect-mode**: *none* },
   body: { **component**: *type-relation*,
   type: { **component**: *object-type*, **attributes**: [{**shape**: *circle*}, {**color**: (!) *yellow*}], **incorrect-mode**: *color* }
   }}

   (b) Guarantee well-defined incorrectness via use of ternary logic

4. Realise caption model as natural language statement (section 4.3)

   (a) Map caption components to DMRS graph snippets, specified as JSON lookup table

   {**shape**: *circle*}: `[type]:_circle_n_1 x?[pers=3] <-- [quant]:default_q`
   {**color**: *yellow*}: `[attr]:_yellow_a_1 e? =1=> [type]:node <-- [quant]:default_q`
   {**component**: *object-type*}: `[type]:_shape_n_1 x?[pers=3] <-- [quant]:default_q`
   {**component**: *type-relation*}: `[rel]:_be_v_id e? -2-> [type]:node`
   {**component**: *existential*}: `[quant]:_a_q --> [rstr]:pred x?[num=sg] <-1- [body]:node`

   (b) Compose snippets to fully specified DMRS graph

```
_a_q --> [rstr]:_circle_n_1 x?[pers=3] <-1- [body]:_be_v_id e? -2-> [type]:_circle_n_1 x?[pers=3]
_a_q --> :type <-- _yellow_a_1 e?
```

(c) Apply post-processing/paraphrase rewriting rules

(d) Pass resulting (D)MRS to ACE plus appropriate grammar (ERG), to obtain the corresponding natural language caption

"A circle is a yellow circle."

5. Output image, correct/incorrect natural language caption, plus their agreement value as given by construction

"A circle is a yellow circle." ⇒ false

4.5 Additional features of simulator architecture

In the following, three additional features of ShapeWorld's system design are discussed: the ability to reverse the realisation process and turn it into a parse-and-verify caption functionality, the relative ease of producing multilingual data, and the potential of intermediate representations in addition to the fully symbolic internal representation on the one hand, and the 'fully natural' language output on the other. It is argued that these are more 'by-products' than additional features in the sense that they were not explicitly implemented but resulted from ShapeWorld's general and principled approach to grounded natural language generation.

Parsing functionality. An advantage of the MRS formalism and bi-directional grammars like the ERG is that the conversion does not just work from (D)MRS to natural language, but also in the – more commonly used – other direction. The ability to parse statements and analyse their formal agreement with a ShapeWorld image consequently comes almost for free, simply by reversing the mapping from caption components to DMRS graph snippets as well as the composition and paraphrasing processing steps.

Parsing a natural language sentence to a caption representation works as follows: first, ACE and the ERG yield a list of possible (D)MRS parses. Going through these, the paraphrase rewriting rules are applied with swapped source and target DMRS subgraph pattern. Subsequently, a top-down construction of the resulting DMRS graph using the available DMRS snippets is attempted, which mirrors the bottom-up composition when producing a DMRS representation of a caption. In other words, the parsing procedure starts by checking whether any of the proposition-level snippets match parts of the target graph using the matching algorithm discussed in section 4.3.2. Whenever successful, the process continues with the proposition's child components, and so on. At each step it is additionally ensured that the entire DMRS graph reconstructed so far still matches the target graph – not to, for instance, mistakenly parse a subject noun phrase and match it as object phrase of a previously identified relation component.

Most of these attempts fail quickly, and only at most one reconstruction attempt will eventually succeed to cover the entire target DMRS graph. By keeping track of the caption components associated with the graph snippets used in this process, the corresponding caption model for the natural language input is recovered.

Being able to assess whether a caption actually applies to an image may have a variety of applications. One application presented later in section 6.3 is to analyse image captioning systems. Part of the reason why the evaluation of generative models is notoriously difficult is the inability to automatically check whether their output is accurate. Even for artificial data, this is generally non-trivial as language gets more complex. While a correct caption alone does not make a 'good' caption for an image, it is arguably a necessary requirement for descriptive captions. Following the principles of unit-testing, the parsing functionality makes it possible to verify that captioning systems satisfy this basic requirement.

Multilingual data. While the current implementation of the ShapeWorld system uses English as the default language, most of the architecture is language-agnostic. A great advantage of using a compositional grammar formalism like (D)MRS is that the semantic graph composition is, in principle, also language-agnostic. Consequently, the only language-specific modules of the system are, on the one hand, the bi-directional MRS-based grammar (ERG for English) that is used in combination with ACE to transform (D)MRS representations into corresponding natural language statements (step 4d in section 4.4) and, on the other hand, the mapping of caption components to DMRS graph snippets according to the underlying grammar (step 4a in section 4.4). The DELPH-IN consortium develops and makes available grammars for a wide range of languages, from more established grammars like JACY for Japanese (Siegel et al., 2016) or Zhong for Chinese (Fan et al., 2015), to many smaller grammars, in some cases for extremely low-resourced languages, in the context of the Grammar Matrix project (Bender et al., 2002). ShapeWorld's abstract domain makes it easy for grammars to be integrated, as only a small vocabulary and relatively basic syntactic constructions are required.

The ability to easily support multilingual data generation is desirable for at least two reasons. First, the fact that the vast majority of available language data, including evaluation benchmarks for deep learning, is English is unsatisfying, as it leaves other languages under-represented. Second, the English language with its rich lexicon, but impoverished inflectional morphology, may exhibit characteristics which are comparatively straightforward for deep learning methods to cope with, and/or methods have mainly been developed which are well-suited to handle these characteristics well (considering, for instance, the many bag-of-words/embeddings-style baselines mentioned in section 2.1). Being able to evaluate models on a range of diverse languages would bring us closer to evaluating the human-like ability of learning to understand language in general. I want to emphasise here that currently dominant methods for obtaining/generating language data, like crowdsourcing, are ill-suited to support multiple languages, as

they essentially require the entire process to be repeated for a new language (see, for instance, the multilingual VQA dataset of Gao et al. (2015)). Even template-based language generation – typical for producing data similar to ShapeWorld – will likely struggle with morphologically richer languages. In contrast, by making use of the independent work of grammar engineers (Bender et al., 2002), DMRS-based generation only requires the minimal effort of translating the 'lexicon' of a system, that is, the mapping of atomic caption components.

As a first proof-of-concept that another language can be integrated straightforwardly, I ported the patterns necessary to generate simple existential statements to the Chinese MRS-based grammar Zhong (Fan et al., 2015), with the help of Huiyuan Xie. Indeed, by just adding the lookup table and the compiled grammar[6], it is possible to produce ShapeWorld data for Chinese, as illustrated in the figure to the right.

Symbolic to surface representations. Both the simulator and the captioner module generate symbolic representations of images and captions, respectively. These can be useful for analysing statistics about the content of the generated data, for sub-selecting certain instance types if the system's configurability does not support specialised setups, or for investigating and comparing model performance with baseline/ceiling systems which process symbolic representations instead of their surface version, and thus simulate 'loss-less' optimal parsing capability of the raw input. Importantly, the symbolic representations on the one hand, and the surface output as image and natural language sentence on the other, are just two 'extremes' of different fully specified representations of the content. Intermediate and/or simplified alternative representations are possible, and can be useful particularly for the latter use case of evaluating a deep learning model using partially processed structured inputs, thus enabling faithful evaluation of later processing steps without being constrained by the quality of preceding steps.

For instance, an image can be represented as a bag-of-objects instead, removing the subtask of identifying objects and their attributes. Alternatively, a caption can be turned into a dependency tree or another linguistic representation as opposed to a natural language sentence, assuming perfect parsing capabilities of the system, to focus on assessing the quality of subsequent processing steps which may involve, for example, TreeRNNs. ShapeWorld currently supports an alternative representation of captions as sequence of symbols representing their caption tree in normal or reverse Polish notation. This intermediate structure can be used, for instance, to train systems that learn to dynamically assemble module networks, like the systems of Hu et al. (2017) and Johnson et al. (2017b) in the context of CLEVR[7].

[6] GitHub project: `https://github.com/delph-in/zhong`.

[7] A version of the PG+EE model (Johnson et al., 2017b) using this intermediate representation can be found as part of my model GitHub repository under `https://github.com/AlexKuhnle/film`.

Below a simple example of a caption in normal and reverse Polish notation. Square brackets signal object-type boundaries for easier reading, but are not actually included in the output.

PN: Existential [ObjectType1 Attribute-shape-pentagon] [Relation-y-rel--1 [ObjectType2 Attribute-color-green Attribute-shape-ellipse]]

RPN: [Attribute-shape-pentagon ObjectType1] [[Attribute-color-green Attribute-shape-ellipse ObjectType2] Relation-y-rel--1] Existential

"A pentagon is above a green ellipse."

Chapter 5

Comparative evaluation of VQA models on ShapeWorld

In the last two chapters, I first introduced the general principles of my evaluation approach, and subsequently the ShapeWorld system which implements a data simulator suitable for detailed investigation of visual question answering models. This chapter presents a comparative experimental analysis of a range of VQA models based on ShapeWorld data. The purpose of this evaluation is to identify the relative strengths and weaknesses of these models, and where their learning behaviour shows no differences.

The chapter is structured as follows: section 5.1 introduces the visual question answering models which will subsequently be evaluated. The presentation attempts to identify 'generic' parts of their architecture, shared by all models, and what the corresponding papers introduce as their core module. To enable a fair evaluation, hyperparameters of the generic parts are aligned across all models. Next, the models are benchmarked on the CLEVR dataset (Johnson et al., 2017a) in section 5.2, for comparison and as sanity check for their correct working. The CLEVR results are unfortunately inconclusive, since in some cases the original implementation or best-effort replication does not show the expected performance. Interestingly, though, the modified versions with unified hyperparameters tend to do better. Section 5.3 describes the ShapeWorld data used in experiments, with a focus on the different types of investigated caption patterns, like existential, relational or number statements. Section 5.4 then presents a range of experimental results, including detailed learning curves over the course of training and a breakdown of how performance of the different models compares per dataset type. Furthermore, section 5.5 investigates targeted architecture modifications for the case of spatial relations, which showed some of the most interesting differences in learning behaviour. Section 5.6 concludes the chapter and discusses how best to interpret the experimental results as well as some general implications.

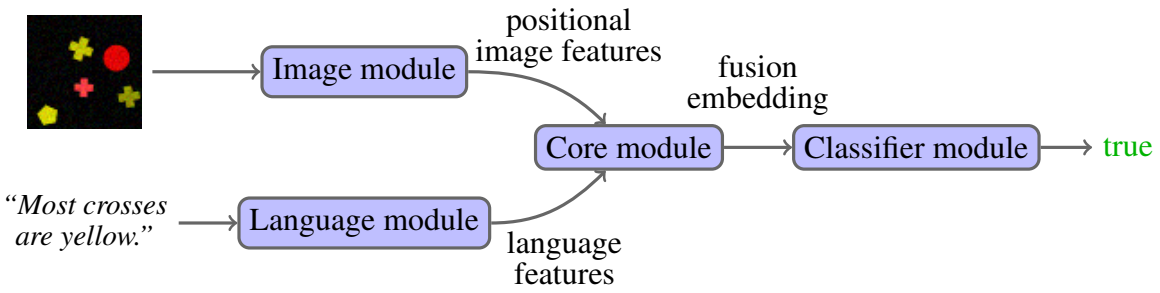


Figure 5.1: The architecture layout shared by all evaluated models. The configuration of image, language and classifier module are unified, so that models only differ in their core module.

5.1 VQA models

This section presents a systematic overview of the VQA models used in later experiments. Besides the two unimodal vision-/language-only baselines, and the multimodal CNN+LSTM baseline, these include: the stacked attention network (Yang et al., 2016), the relation network (Santoro et al., 2017), the early-fusion multimodal core module (Malinowski and Doersch, 2018), and the FiLM model (Perez et al., 2018). As illustrated in figure 5.1, the presentation tries to disentangle the generic image and language feature extraction as well as the answer classification modules from the essential part of the respective architecture, where the models actually differ in their method to combine image and language information. To enable a fair evaluation, my model implementations largely use the same non-core module configurations.

Why not use the model configurations from the original papers? Minor architectural differences can have a big impact on how well a model is able to learn a task. For instance, section 2.1 reported a range of cases where hyperparameter tuning alone substantially improved performance of a baseline model, including the CNN+LSTM model (Lu et al., 2015) and the stacked attention network (Santoro et al., 2017) in the context of visual question answering. While all of the aforementioned models are characterised by a core module and otherwise follow the same architecture layout (see figure 5.1), they nonetheless exhibit other minor differences, which may be the result of tuning or due to the fact that more recent models leverage previous improvements. Since the experiments in this chapter are supposed to compare the strengths and weaknesses of these models, unifying the configuration of the generic modules prevents spurious differences such as differing capacity of the language module or the application of batch normalisation in the visual module from affecting the results, thus a more faithful comparison of the core modules, even if the changes should result in an inferior model variant (which does not seem to be the case, according to section 5.2). In summary, the unified setup trades off optimal absolute performance in favour of improved comparability. The comparison between original and unified setup on the CLEVR dataset in section 5.2 confirms the validity of these concerns, as does the differing results for the original FiLM model on ShapeWorld data in chapter 6 (where the evaluation is focused only on FiLM and thus comparability is not important).

In the following, section 5.1.1 specifies the unified hyperparameter variant of the models, and section 5.1.2 points out the differences to their original version.

5.1.1 Unified hyperparameter setup

Figure 5.1 illustrates the basic architecture layout shared by all of the evaluated models: (a) an image module processes the (raw or pretrained) input image features and outputs a map of positional image embeddings; (b) a language module embeds and processes the input question/caption words and outputs a language embedding; (c) a core module combines positional image and language features and outputs a fusion embedding; and (d) a classifier module maps the fusion vector to a softmax distribution over responses. In the following, each module is described in detail, including layer sizes and other hyperparameter choices.

Image module. The visual input is processed by a sequence of convolutional layers. Each layer consists of a convolution operation (LeCun et al., 1989) with kernel size 3×3 and stride size 1 or 2, followed by two-rank batch normalisation (Ioffe and Szegedy, 2015) and subsequently a rectified linear unit (Glorot et al., 2011). The output image features are of size $w \times h \times c$, that is, $w \cdot h$ positional image embeddings of c dimensions. The number of convolutional layers, number of kernels and stride size per layer – and consequently the size of the image features – depends on data and architecture:

- CLEVR with pretrained ResNet-101 image features: Following (Johnson et al., 2017a; Johnson et al., 2017b; Perez et al., 2018), after resizing to images of size $224 \times 224 \times 3$, features of size $14 \times 14 \times 1024$ are extracted from the conv4 layer of a pretrained ResNet-101 (He et al., 2016). In this case, the image module consists of only one convolutional layer with 128 kernels and stride 1, yielding output features of size $14 \times 14 \times 128$. The CNN+LSTM+REL model uses a stride size of 2 instead, yielding an output of size $7 \times 7 \times 128$ (to keep the number of pairwise combinations moderate).

- CLEVR from raw images: Similar to (Santoro et al., 2017; Malinowski and Doersch, 2018; Perez et al., 2018), but without resizing images, four convolutional layers with 128 kernels and stride 2 are applied to the input, thus reducing the initial size of $320 \times 240 \times 3$ to $30 \times 20 \times 128$. The CNN+LSTM+REL model adds an additional convolutional layer, yielding an output of $15 \times 10 \times 128$ (again, to keep the number of pairwise combinations moderate).

- ShapeWorld from raw images: The same configuration as above is used, but with only three instead of four convolutional layers, due to the smaller input image dimensions of $64 \times 64 \times 3$. The output image features are thus of size $8 \times 8 \times 128$.

Language module. The words of the input question/caption are mapped to 128-dimensional word embeddings and processed by an LSTM (Hochreiter and Schmidhuber, 1997) – or GRU (Cho et al., 2014) in case of CNN+GRU+FILM – of size 512, with the final processed word as the 512-dimensional language embedding output. The CNN+LSTM+REL model uses an LSTM of size 128 instead (to keep the size of the pairwise combination embeddings moderate).

Core modules.

- CNN baseline: No language input. A 128-dimensional linear transformation is applied to the 128-dimensional positional image embeddings, with subsequent max-pooling over all embeddings to obtain a 128-dimensional output fusion embedding.
- LSTM baseline: No image input. The 512-dimensional language feature vector is passed on as the output fusion embedding.
- CNN+LSTM baseline, combination of the two modules above, which combines visual and language features after global pooling: a linear transformation is applied to the 128-dimensional positional image embeddings before max-pooling, and the resulting vector is concatenated with the 512-dimensional language embedding to yield the 640-dimensional output fusion embedding.
- CNN+LSTM+SA model (stacked attention, Yang et al. (2016)), which infuses the language features before global pooling by conditioning a series of attention maps over image features: An initial 256-dimensional linear transformation is applied to the 128-dimensional positional image embeddings and the 512-dimensional language embedding, respectively. Subsequently, two stacked attention layers process the input. For each layer, a 256-dimensional linear transformation processes the positional image and language embeddings, respectively, before adding them and applying a tanh activation function. Another 1-dimensional linear transformation with subsequent softmax operation gives the multiplicative attention map over positional image embeddings. The resulting 256-dimensional weighted sum of positional image embeddings is added to either the transformed input language embedding or the output of the previous layer. This yields a final 256-dimensional output fusion embedding.
- CNN+LSTM+REL model (relation module, Santoro et al. (2017)), which combines pairs of positional image with language features and processes them in a series of additional fully-connected layers before global pooling: An initial 32-dimensional linear transformation turns the 128-dimensional positional image embeddings into 32-dimensional image features, which then are concatenated with a 2-dimensional map of relative spatial coordinates. Subsequently, each pair of 34-dimensional embeddings plus a copy of the 128-dimensional language embedding are concatenated, and processed by four 256-dimensional fully-connected layers with

rectified linear units. The resulting $(w \cdot h)^2$ 256-dimensional vectors are sum-pooled to obtain a single 256-dimensional output fusion embedding.

- CNN+LSTM+MC model (multimodal core, Malinowski and Doersch (2018)), which combines visual and language features and processes them in a series of additional fully-connected layers before global pooling: Each 128-dimensional positional image embedding is concatenated with a copy of the 512-dimensional language embedding. Following two-rank batch normalisation, each positional vector is processed by four 256-dimensional fully-connected layers with rectified linear units. Finally, sum-pooling the resulting 256-dimensional vectors yields a single 256-dimensional output fusion embedding.

- CNN+GRU+FILM model (feature-wise linear modulation, Perez et al. (2018)), which infuses the language features before global pooling by conditioning the modulation values following batch normalisation in a series of additional convolutional layers: The image features are processed by four FiLM layers. For each layer, the 128-dimensional positional input embeddings are concatenated with a 2-dimensional map of relative spatial coordinates and processed by a 128-dimensional fully-connected layer with rectified linear unit. Subsequently, a convolution operation with 128 kernels of size 3×3 and stride size 1 is applied, followed by two-rank batch normalisation, however, instead of learned scale and offset values, these are obtained via two 128-dimensional linear transformations from the 512-dimensional language embedding. A rectified linear unit is applied to the output before being added as residual to the vectors before the convolution operation. Finally, the 128-dimensional positional vectors of the fourth FiLM layer are, again, concatenated with a spatial coordinate map and processed by a final 128-dimensional linear transformation, followed by two-rank batch normalisation, rectified linear unit, and then max-pooled to a 128-dimensional output fusion embedding.

Classifier module. The 128-, 256- or 640-dimensional output fusion embedding of the core module is processed by a fully-connected layer of size 1024, followed by one-rank batch normalisation and a rectified linear unit, before being mapped to answer logits by another linear layer and passed through a softmax operation to retrieve a distribution over answers.

Optimisation. Models are trained by Adam (Kingma and Ba, 2015), a first-order gradient-based stochastic optimiser, with a learning rate of $3 \cdot 10^{-4}$ and mini-batches of size 64.

Codebase and contribution. My implementation(s) can be found as part of the GitHub repository under `https://github.com/AlexKuhnle/film`, which extends and modifies the FiLM (Perez et al., 2018) repository under `https://github.com/ethanjperez/film`, which itself is based on the original PG+EE repository under `https://github.com/facebookresearch/clevr-iep` (Johnson et al., 2017b). Besides modifying the existing

code to support the unified hyperparameter setup and to make it compatible with ShapeWorld, I added the implementation of the CNN+LSTM+REL and CNN+LSTM+MC model.

Effective differences between unified models. The core modules differ in a variety of aspects. First, all models but the CNN+LSTM baseline rely on early as opposed to late fusion, that is, they combine visual and language information before pooling all positional embeddings into a single fusion embedding. In case of early fusion, the fusion mechanism is applied either pointwise to, or pairwise between all positional embeddings. Language and (pairs of) positional image embeddings are combined either by concatenation, pointwise addition or an affine operation (multiplication plus addition). Some core modules add a map of relative spatial coordinates to the positional image embeddings before processing them. The core module itself consists either of one or more fully-connected layers applied per position, or a residual convolutional layer applied to a local window of embeddings. The entire process may be repeated multiple times. Finally, the processed positional embeddings are pooled to a global embedding either via concatenation/flattening, global sum- or max-pooling, or by weighted attention. The following table summarises the differences between the core modules with respect to these key characteristics.

	Multimodal fusion			Coord map	Module operation	Depth	Positional pooling
	When	Type	Operation				
CNN+LSTM	late	–	concat	no	–	1	concat
. . . +SA	early	pointwise	additive	no	fc	2	attention
. . . +REL	early	pairwise	concat	yes	$4 \times$ fc	1	sum
. . . +MC	early	pointwise	concat	no	$4 \times$ fc	1	sum
. . . +FILM	early	pointwise	affine	yes	res conv	4	max

5.1.2 Original model hyperparameters

The unified hyperparameter choice differs from the original models as described in the corresponding paper[1]. The following list points out differences between the original version of the model and my replication.

- CNN+LSTM+SA-ORIG:
  - Image module (CLEVR with pretrained features): two (instead of one) convolutional layers, no batch normalisation.
  - Language module: 300-dimensional (instead of 128-dimensional) word embeddings, two-layer LSTM of size 256 (instead of 512).

[1] In case of the CNN+LSTM+SA model, I base the comparison on the default hyperparameters of the implementation in the PG+EE/FiLM GitHub repository, which presumably is (close to) the version used for the PG+EE, FiLM and original CLEVR paper (Johnson et al., 2017b; Perez et al., 2018; Johnson et al., 2017a).

- Core module: initial 256-dimensional linear transformation only for positional image embeddings (and not for language embedding), 512-dimensional (instead of 256-dimensional) stacked attention linear transformations, additional final 512-dimensional linear transformation and subsequent 2×2 max-pooling.
- Classifier module: no batch normalisation.
- Optimisation: learning rate of $5 \cdot 10^{-4}$ (instead of $3 \cdot 10^{-4}$).

- CNN+LSTM+REL-ORIG:
  - Image module (CLEVR from raw images): 24 (instead of 128) kernels per convolutional layer. Not replicated: images are resized to size 128×128 (instead of 320×240) and processed by four (instead of five) convolutional layers, resulting in image features of size 8×8 (instead of 15×10).
  - Language module: 32-dimensional (instead of 128-dimensional) word embeddings.
  - Core module: no initial 32-dimensional linear transformation for positional image embeddings (which are 24-dimensional and thus already small here).
  - Classifier module: two 256-dimensional (instead of one 1024-dimensional) layers, no batch normalisation, second layer additionally with a dropout rate of 0.5.
  - Optimisation: learning rate of $2.5 \cdot 10^{-4}$ (instead of $3 \cdot 10^{-4}$).

- CNN+LSTM+MC-ORIG:
  - Image module (CLEVR from raw images), not replicated: images are resized to size 256×256 (instead of 320×240), resulting in image features of size 16×16 (instead of 30×15).
  - Language module: dimensions of word embeddings not clearly specified, I use 64-dimensional (instead of 128-dimensional) word embeddings here, LSTM of size 128 (instead of 512).
  - Core module: same.
  - Classifier module: not clearly specified, I use no batch normalisation and a dropout rate of 0.5 here.
  - Optimisation: same.

- CNN+GRU+FILM-ORIG:
  - Image module (CLEVR from pretrained features): same.
  - Image module (CLEVR from raw images): convolutional layer with kernel size 4×4 (instead of 3×3). Not replicated: images are resized to size 224×224 (instead of 320×240), resulting in image features of size 14×14 (instead of 30×15).

- Language module: 200-dimensional (instead of 128-dimensional) word embeddings, GRU of size 4096 (instead of 512).
- Core module: final 512-dimensional (instead of 128-dimensional) linear transformation.
- Classifier module: same.
- Optimisation: additional weight decay of 10^{-5}.

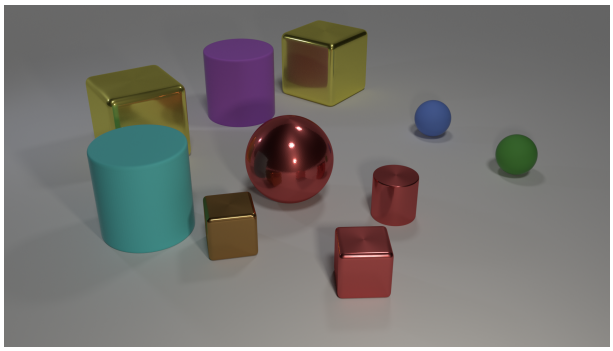
5.2 Experiments on the CLEVR dataset

The CLEVR dataset (Johnson et al., 2017a) has had a big impact on research around visual question answering and inspired a range of new models, including most of the ones presented in section 5.1. Moreover, CLEVR consists of abstract data and is supposed to serve for diagnostic evaluation purposes – two aspects which make it similar to the nature of data and motivation behind the ShapeWorld framework. It is thus a natural starting point for the experimental part to assess model performance on CLEVR.

Previous research has mostly followed the precedent of Johnson et al. (2017a) and trained models on visual features extracted from a pretrained ResNet model. Only in some cases, a simple feature extractor for raw images was trained as part of the architecture (Santoro et al., 2017; Malinowski and Doersch, 2018; Perez et al., 2018). I conduct experiments using both pretrained and raw image features, and also compare my implementation of a unified hyperparameter version with the original model configurations. The purpose is mainly to connect later results on ShapeWorld – which will not use a pretrained feature extractor and only focus on the unified version – with results in the literature, and to present a complete overview of the different experimental variants.

5.2.1 Data

The CLEVR visual question answering dataset consists of rendered images of abstract three-dimensional scenes, associated with questions and their ground-truth answer which are generated from a variety of templates. Figure 5.2 shows an example instance. Similar to ShapeWorld, CLEVR's internal world models are defined by a list of objects located on a two-dimensional plane, although rendered in three dimensions. Objects have one of three shape types (*"cube"*, *"sphere"* or *"cylinder"*), two discrete sizes (*"small"* or *"large"*), two materials (*"shiny metal"* or *"matte rubber"*), and eight colours. Questions are categorised into different types, depending on the required ability to correctly answer them: existential or counting questions, questions asking to compare object numbers (*"equal"*, *"less"* or *"more"*), or questions either querying or asking to compare the attribute of an object (*"shape"*, *"size"*, *"material"* or *"colour"*). Overall, there are 28 answers, of which a subset are applicable depending on the question type: yes/no, numbers from 0 to 10, and the 15 attribute values.



- How many small spheres are there? – 2
- What number of cubes are small things or red metal objects? – 2
- Does the metal sphere have the same colour as the metal cylinder? – Yes
- Are there more small cylinders than metal things? – No
- There is a cylinder that is on the right side of the large yellow object behind the blue ball; is there a shiny cube in front of it? – Yes

Figure 5.2: An image and five example questions plus corresponding answers from the CLEVR dataset.

The CLEVR training set consists of 70,000 images with 10 questions each, thus overall 700k training instances. In the following, only the number of training iterations is reported – given the batch size of 64 used in all experiments, 100k iterations are equivalent to roughly 9.1 epochs. The validation set contains another 15,000 images with 10 questions each, summing up to 150k validation instances. Accuracy is always measured on the entire validation set, every 2,000 iterations for the first 10k and every 5,000 iterations afterwards.

5.2.2 Results

Performance of baseline models. The vision-only CNN baseline achieves an accuracy of slightly above 20% (see figure 5.3). The language-only LSTM baseline reaches around 47% accuracy in accordance with Johnson et al. (2017a), and learning already plateaus after only 5-10k iterations. The multimodal CNN+LSTM baseline obtains around 56% when using pretrained image features, saturating after roughly 50k iterations, and 58% when learning from raw images, plateauing after 80-100k iterations. This is slightly better than the 52.3% reported by Johnson et al. (2017a) and subsequent papers.

Performance of original models. The learning curve for the CNN+GRU+FILM model reaches around 96% accuracy after 200k iterations using pretrained image features (see figure 5.3), and the same score for raw images after 300k iterations, without indicating saturation in either case, which corresponds to the results reported by Perez et al. (2018). The CNN+LSTM+MC model with pretrained features obtains around 86% accuracy in the same time and, surprisingly, only around 71% when using raw images. These accuracy levels diverge from the better results of Malinowski and Doersch (2018), and may either be simply a matter of training the models for longer, or due to insufficient details on implementation and hyperparameters in their paper. Performance of the CNN+LSTM+SA model using pretrained image features stays below the CNN+LSTM baseline, in stark contrast to the 68.5% of Johnson et al. (2017a) despite the fact

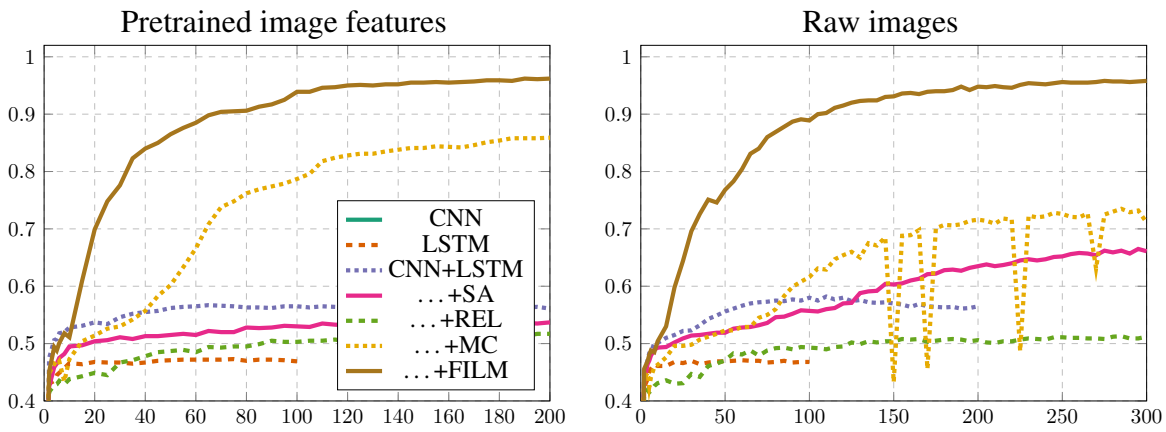


Figure 5.3: Performance of original models on the CLEVR dataset (*x-axis*: iterations in 1000, *y-axis*: accuracy).

that my codebase is based on theirs (to be precise: the one of Johnson et al. (2017b)) with default hyperparameters. Interestingly, CNN+LSTM+SA obtains 66% accuracy after 300k iterations for raw images, which fits better with the results of Johnson et al. (2017a) for pretrained features. Finally, the CNN+LSTM+REL model does not improve upon the CNN+LSTM baseline in either case, which is not consistent with the results of Santoro et al. (2017).

Performance of models with unified hyperparameters. The learning curve for most models changes substantially when moving to the unified configuration (see figure 5.4). For the CNN+LSTM+MC model, performance stays almost the same – unsurprisingly, as its architecture changes comparatively little in the unified setup. The CNN+GRU+FILM model is the only one whose performance decreases substantially, reaching only 88% after 200k iterations with pretrained features, and 84% after 300k using raw images. The other two models both improve upon the CNN+LSTM baseline in this setup. In the case of pretrained image features, all models are roughly on par, with accuracies between 83-88%, whereas performance levels vary between 68%-84% using raw images. The 74% accuracy of the CNN+LSTM+SA model is similar to the 76.6% of the (unpublished) implementation of Santoro et al. (2017), which is *"trained fully end-to-end"*, so presumably learned from raw images. Interestingly, accuracy of the same model with pretrained features is much better than any reported result for this model.

5.2.3 Conclusion

Unfortunately, only some of the experimental results on CLEVR are in accordance with the literature. While only the CNN+GRU+FILM model reaches roughly the expected accuracy level in all cases, the unified hyperparameter setting makes a big difference for CNN+LSTM+REL and CNN+LSTM+SA. This may be due to the use of batch normalisation.

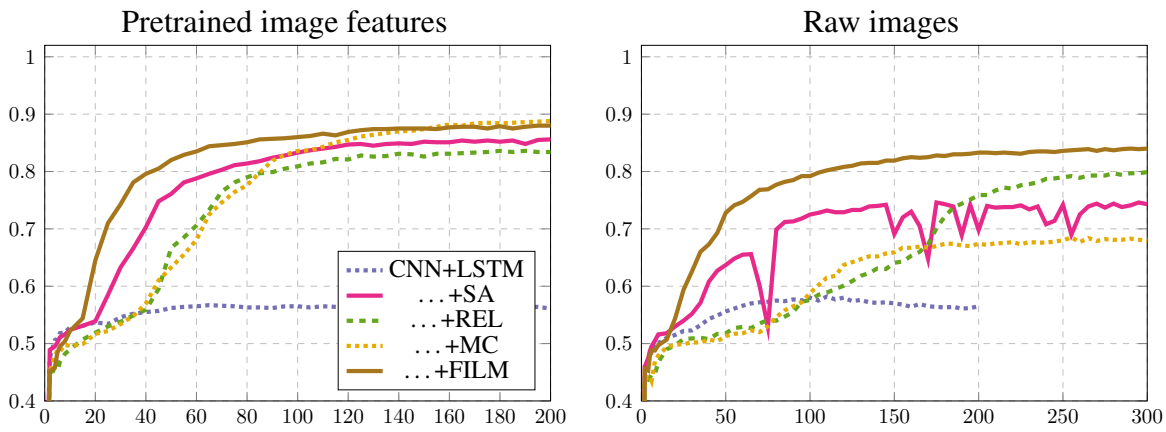


Figure 5.4: Performance of models with unified hyperparameters on the CLEVR dataset (*x-axis*: iterations in 1000, *y-axis*: accuracy).

It is well-known that even simple reproduction of machine learning results can be problematic, as discussed in section 2.1. The situation here is aggravated by the fact that open-source code was not available for all of the evaluated models, and in some cases details of the architecture were not sufficiently specified in the corresponding paper. However, since the aim of this section is not to tune models for optimal performance on CLEVR, but just to compare my implementations, these issues are not further investigated. Importantly, though, the results generally confirm that the unified model variants obtain good results: they learn to handle CLEVR instances substantially better than the baselines, and often even better than my implementation of their original version.

5.3 ShapeWorld datasets

This section describes the various configurations for ShapeWorld data which form the basis for later experiments. Each configuration focuses on one type of caption pattern like, for instance, statements containing spatial relations. The following overview categorises the different types of datasets (see section 4.2.1 for the corresponding caption components and their interpretation).

EXISTENTIAL	SINGLE-EXISTENTIAL	EXISTENTIAL ONE/TWO/THREE-SHAPE	
		EXISTENTIAL FULL	
	DOUBLE-EXISTENTIAL	RELATIONAL-TRIVIAL	
		LOGICAL	
QUANTIFICATION	NUMBERS		
	QUANTIFIERS		
RELATIONAL	NON-SPATIAL	ATTRIBUTE-EQUALITY	
		ATTRIBUTE-RELATIVE	
	SPATIAL	SPATIAL-EXPLICIT	
		SPATIAL-IMPLICIT	SPATIAL-COMPARATIVE
			SPATIAL-SUPERLATIVE

Since data generation takes much longer than model training for ShapeWorld data, sufficiently big datasets are produced once and reused for all experiments. Training datasets consist of 500k instances. Validation and test sets each consist of additional 10k instances, with validation instances following the same configuration as the training data, while test instances may additionally exhibit withheld numbers of objects and attribute combinations. Scenes generally contain 1/4/5 to 10/15 objects depending on the dataset, to encourage 'interesting' non-trivial situations[2]. For all datasets, 5, 10 and 15 are withheld numbers of objects which are only generated for test data. Special cases of datasets with instances consisting of only one, two or three shapes are indicated by the labels ONE/TWO/THREE-SHAPE, and whether overlapping objects are avoided is indicated by COLLISION-FREE.

The following object attribute combinations are withheld for training and validation data: *"red square"*, *"green triangle"*, *"blue circle"*, *"yellow rectangle"*, *"magenta cross"*, *"cyan ellipse"*, *"red/green/blue/grey pentagon"*, *"grey square/triangle/circle/pentagon"*. The fact that shape and colour attributes appear in multiple combinations and within varying caption patterns encourages systems to disentangle the two properties in a factored attribute-level representation. Test instances and their withheld combinations consequently evaluate whether such a representation is learned, as otherwise it would not be possible to generalise to these unseen object descriptions. In particular *"pentagons"* and *"grey shapes"*, for which only about half of the possible combinations are seen during training, test the degree of robustness of this generalisation ability.

Existential statements. SINGLE-EXISTENTIAL datasets consist of simple statements referring to the existence of (at least) one object of a certain description, which may be partially underspecified, that is, only mention either the shape or colour of an object. The following list illustrates the different possible surface statements referring to a red square in a scene.

- *"There is a square."*
- *"There is a red shape."*
- *"A shape is a square."*
- *"A shape is red."*
- *"There is a red square."*
- *"A shape is a red square."*
- *"A square is red."*
- *"A red shape is a square."*

SINGLE-EXISTENTIAL datasets can be seen as a language-variant of the object recognition vision task. The language representation of the 'object category' incentivises to learn attribute-factored representations as opposed to independent classes, which allow a system to generalise to unseen combinations.

Statements with logical connectives. The LOGICAL dataset combines two existential statements with one of the following logical connectives: *"and"*, *"or"*, *"if"* or *"if and only if"*.

[2] SINGLE-EXISTENTIAL: 1-10 objects; DOUBLE-EXISTENTIAL and RELATIONAL: 4-10 objects; QUANTIFICATION: 5-15 objects.

The existential components each refer to a different object, and either of them may be partially underspecified. The following list contains an example for each connective.

- *"There is a square and a shape is a circle."*
- *"There is a square or a circle is green."*
- *"A square is red if there is a circle."*
- *"A square is red if and only if there is a green circle."*

The LOGICAL dataset requires to detect the existence or non-existence of two independent descriptions of objects. The connective determines which combinations of non-/existence are considered correct. Note that it is not necessary to keep track of both sets of objects simultaneously. For instance, in case of an *"or"* statement, either the second part can be ignored if the first description already applies, or the first can be forgotten if it does not apply. This distinguishes the dataset from the RELATIONAL datasets below.

Statements with numbers or quantifiers. The QUANTIFICATION datasets both consist of quantified statements about a set of objects. In the case of NUMBERS, the quantification is based on an absolute number, whereas QUANTIFIERS statements specify the fraction relative to the total number of objects of a description. In addition, one of the following comparing modifiers defines the quantification more precisely: *"more than"*, *"at least"*, *"exactly"*, *"not"*, *"at most"* or *"less than"*. A variant of the dataset without different modifiers, just *"exactly"*, is indicated by the suffix -EXACT.

The NUMBERS dataset uses numbers from *"zero"* to *"five"*, with one example per number given in the following list[3]

- *"More than zero shapes are squares."*
- *"At least one shape is red."*
- *"Exactly two shapes are red squares."*
- *"Not three squares are red."*
- *"At most four red shapes are squares."*
- *"Less than five shapes are red squares."*

The QUANTIFIERS dataset is based on the fractions *"half"*, *"third"* and *"quarter"*, in addition to the 'trivial' fractions *"no"* and *"all"*. An example for each fraction can be found in the following list.

- *"More than no shape is a square."*
- *"At least a quarter of the shapes is red."*
- *"Exactly a third of the shapes is a red square."*
- *"Not half the squares are red."*
- *"At most two thirds of the red shapes are squares."*
- *"Less than three quarters of the shapes are red."*
- *"Not all red shapes are squares."*

[3]Some of the sentences may sound unnatural to English speakers, however, I decided to treat numbers/quantifiers and modifiers as fully compositional in ShapeWorld. Note that models in this thesis are trained from scratch on the resulting data, but the captioner can be configured to exclude unnatural combinations, for instance, when using pretrained word embeddings or language models.

The crucial difference between NUMBERS and QUANTIFIERS in terms of quantification complexity is that NUMBERS statements can be correctly answered solely by counting the number of objects satisfying the combined description of noun and verb phrase (*"Two squares are red."* $\rightarrow$ *"red squares"*), while QUANTIFIERS statements generally require to compare the cardinality of this object set relative to the number of objects in agreement with only the noun phrase part of the description (*"Half of the squares are red."* $\rightarrow$ *"red squares"* relative to *"squares"*). Note also that the -EXACT version with only one modifier, while less complex in terms of linguistic variety, does not contain approximate modifiers like *"at most"*, and thus requires more precise recognition of numbers.

Relational statements. The dataset category RELATIONAL comprises various relational statements between two or more objects. Where the relation requires an additional comparison object – for instance, *"closer to... than"* – this description is constrained to unambiguously refer to a single object in the scene. RELATIONAL datasets are further distinguished between the type of relation they contain, which are described in the following paragraphs.

First, the RELATIONAL-TRIVIAL dataset consists of 'trivial' statements without relational content beyond the co-existence of two objects.

- *"A square exists besides a green shape."*

The ATTRIBUTE-EQUALITY dataset comprises relational statements which compare the shape or colour of two objects, whether they are the same or different.

- *"A red shape is the same shape as a green shape."*
- *"A red shape is a different shape from a green shape."*
- *"A square is the same colour as a circle."*
- *"A square is a different colour from a circle."*

These instances do not mention the shape/colour in question, as otherwise they would effectively reduce to a kind of existential statement: for instance, *"A red shape is the same shape as a green square."* reduces to *"There is a red square."*, since only the shape information in *"green square"* is relevant for the first part of the sentence.

The ATTRIBUTE-RELATIVE dataset contains relations comparing either the size of a shape or the shade of a colour of two objects. Note that these relations implicitly require the same shape/colour to avoid ambiguous comparisons, which is why the corresponding attribute is only mentioned once.

- *"A red shape is smaller than a green circle."*
- *"A square is darker than a green circle."*
- *"A red shape is bigger than a green circle."*
- *"A square is lighter than a green circle."*

The SPATIAL-EXPLICIT dataset involves various spatial relations, including two relying on a third comparison object for relative distances. The two relations *"behind"* and *"in front of"*, which require overlapping objects, are excluded in the case of COLLISION-FREE datasets.

- *"A square is to the left of a circle."*
- *"A red square is above a circle."*
- *"A red square is behind a circle."*
- *"A square is closer to the triangle than a circle."*
- *"A square is to the right of a green circle."*
- *"A red square is below a green circle."*
- *"A red square is in front of a green circle."*
- *"A square is farther from the triangle than a green circle."*

Besides these 'explicitly' relational statements, ShapeWorld supports two other forms of implicit spatial statements, which consist of an adjectival form of one of the relations above. The SPATIAL-COMPARATIVE dataset comprises statements with adjectives in positive/comparative form. They require the object set described by the noun phrase to contain exactly two objects, between which the spatial relation selects the referred target.

- *"The left square is red."*
- *"The right red shape is a square."*
- *"The upper circle is green."*
- *"The lower green shape is a circle."*
- *"The red shape closer to the triangle is a square."*
- *"The square farther from the triangle is red."*

The SPATIAL-SUPERLATIVE dataset consists of similar statements with adjectives in superlative form. Here, the noun phrase refers to at least two objects, of which the one 'maximally' satisfying the spatial relation – that is, all pairwise comparisons with the other objects under consideration – is selected.

- *"The leftmost square is red."*
- *"The rightmost red shape is a square."*
- *"The uppermost circle is green."*
- *"The lowermost green shape is a circle."*
- *"The red shape closest to the triangle is a square."*
- *"The square farthest from the triangle is red."*

5.4 Experiments

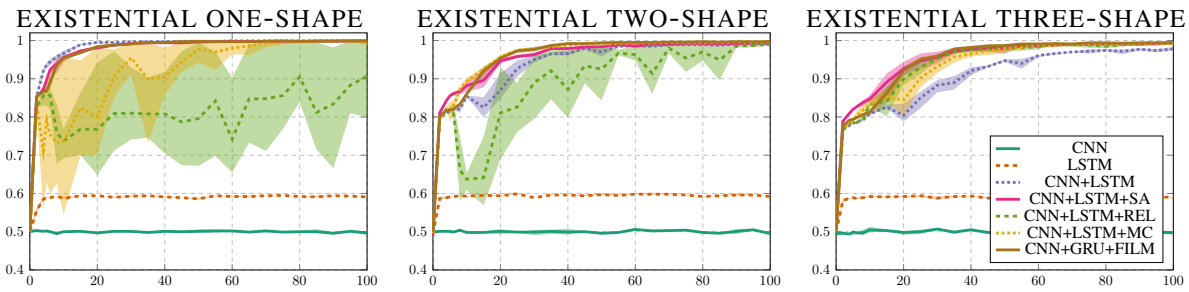
In the following, the experimental results of evaluating the visual question answering models on ShapeWorld data are discussed. Every experiment is run three times and accuracy mentions generally refer to the average over these three runs. Plots of learning curves here and in the remainder of this chapter indicate iterations in 1,000 on the x-axis, accuracy on the y-axis, and additionally minimum and maximum performance amongst the three runs as shaded area. Models are trained for 100-200k iterations on the respective dataset. For comparison, 100k iterations roughly correspond to 13 epochs over the 500k instances dataset, given the batch size of 64. Accuracy is measured on the entire validation set, every 2,000 iterations for the first 10k and every 5,000 iterations afterwards.

Single-modality baselines. The vision-only CNN baseline performs at chance level most of the time, reaching a maximum of 55% on the RELATIONAL-COMPARATIVE dataset, which suggests a minor visual bias (baseline performance curves included in plots throughout the section). Of the agreeing instances in this dataset, the CNN model gets 50% right, while its accuracy is 65% on disagreeing instances. Consequently, the bias seems to be due to certain object configurations making an agreeing caption less likely, so that based on the image alone an instance can be judged as more likely incorrect.

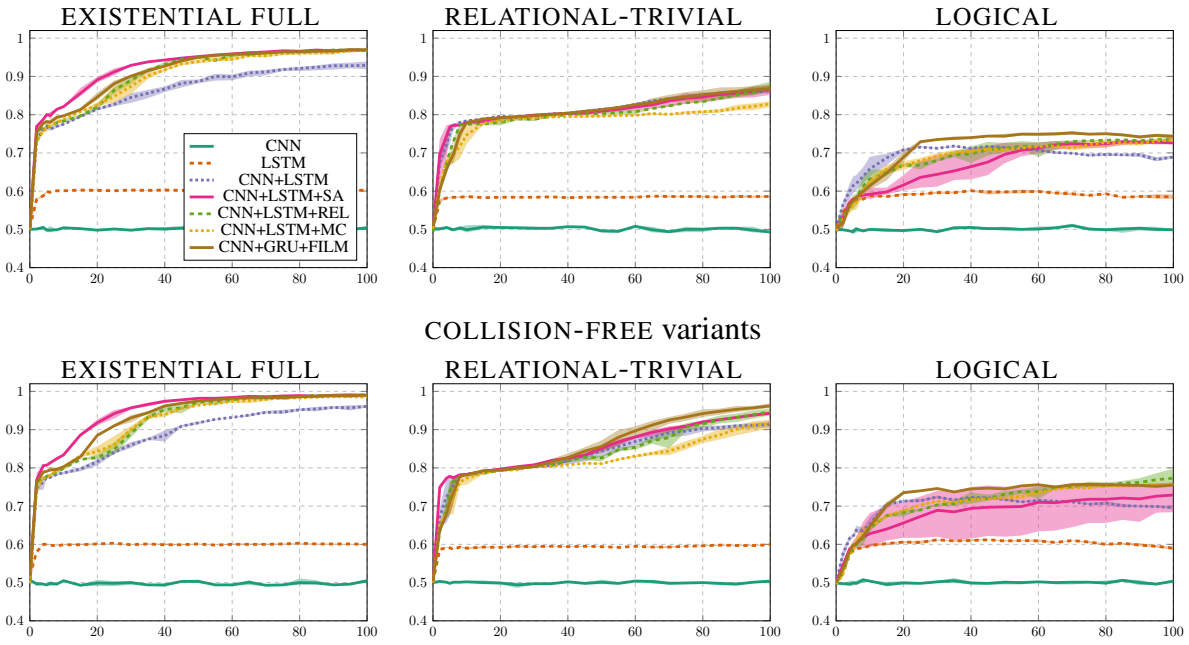
The language-only LSTM baseline achieves an accuracy of around 60% on most datasets. Analysing the detailed breakdown of its performance reveals that the LSTM model learns to consistently classify statements referring to a test combination as incorrect. The language bias is thus primarily caused by the withheld test attribute combinations, which can occur as part of the training data when an attribute or object-type caption component is invalidated. For instance, they account for 7.5% of the validation data in the EXISTENTIAL FULL dataset, which the LSTM model gets always correct. This effect is amplified in the case of the LOGICAL dataset due to two existential sub-statements, where 14.5% of its instances mention a test combination, and the model identifies them correctly 88% of the time. For the other datasets, occurrences of test combinations are generally less likely, since the captions are often invalidated by, for instance, changing the relation instead of the attributes of an object description. Apart from LOGICAL, the language bias does not exceed 62% and is often below 60%.

Overall, considering the 50% chance level, a visual bias of at most 55% and a language bias of at most 62% (except for LOGICAL) confirms that the data generation procedure with its bias-reducing mechanisms largely succeeds in preventing undesired cues in the data.

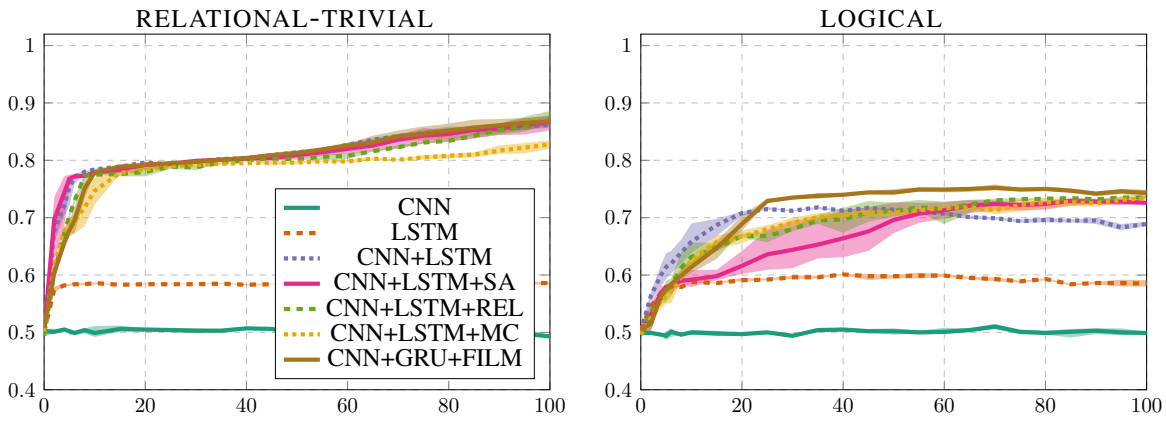
Very simple datasets can cause learning instability. The EXISTENTIAL ONE-SHAPE dataset is intended as a 'sanity check' that every non-trivial model is expected to learn perfectly. While this is indeed the case for most models after around 20-30k iterations, surprisingly, the learning process of the CNN+LSTM+MC and especially the CNN+LSTM+REL model are unstable. CNN+LSTM+MC converges after around 70k iterations, and CNN+LSTM+REL does not reliably manage to solve the dataset within 100k iterations. Note that the high degree of fluctuation between runs observed here, as indicated by the shaded area, is rarely seen for other experiments. Since learning in other cases is more robust, this suggests that the two models do not cope well with the trivial simplicity of the ONE-SHAPE data. Indeed, when moving to the TWO-SHAPE variant of EXISTENTIAL data, the training process of the CNN+LSTM+MC model stabilises and the variance of CNN+LSTM+REL reduces markedly reaching 99% after 90k iterations. Finally, the learning curves on the THREE-SHAPE variant look like the ones for the unmodified EXISTENTIAL FULL dataset.



Partially occluded objects affect performance negatively. Comparing the performance of models on the EXISTENTIAL FULL dataset and its COLLISION-FREE variant, ∼97% versus ∼99% accuracy, suggests that visually overlapping objects have a small negative impact on learnability of the dataset. This effect is observed with varying degree for all datasets, whenever it is investigated. For instance, the difference on RELATIONAL-TRIVIAL is 7-9% for most models, whereas on LOGICAL it is only at most 3-4%. Results in the following are always for datasets with potentially overlapping objects.



Two object references are substantially harder. All models learn to solve the EXISTENTIAL FULL dataset almost perfectly at 97% accuracy after 100k iterations, with only the CNN+LSTM baseline lacking behind by 4-5%. However, performance drops substantially on the DOUBLE-EXISTENTIAL datasets, which combine two existential predications in one statement. All models except the unimodal baselines show virtually the same performance on the RELATIONAL-TRIVIAL dataset, reaching 78% accuracy within the first 10k iterations and then slowly advancing to 87% after 100k iterations. Accuracy levels decrease further for LOGICAL, in total by almost 25%, to 69% for the CNN+LSTM baseline and 73-74% else.



Further investigation of this observation reveals that, generally, object descriptions mentioning the shape are not learned as well as ones specifying the colour of an object. Even in case of EXISTENTIAL FULL, colour-only captions are always handled correctly, while shape-only with only 95-96% accuracy, or 90% in the case of CNN+LSTM. This gap widens for the two DOUBLE-EXISTENTIAL datasets. As can be seen in table 5.5, colour information consistently improves accuracy (s-only $<$ 2s&c / s&c $<$ s&2c), whereas shape mentions actually hurt performance (c-only $>$ s&2c / s&c $>$ 2s&2c, however, the latter only for LOGICAL). Note that this is not due to the data distribution, which is generally symmetric between shape and colour.

	c-only	s-only	s&c	s&2c	2s&c	2s&2c
EXISTENTIAL ONE-SHAPE	100% (20%)	100% (20%)	100% (60%)			
EXISTENTIAL FULL	100% (20%)	95% (20%)	96% (60%)			
EXISTENTIAL COLLISION-FREE	100% (20%)	98% (20%)	99% (60%)			
RELATIONAL-TRIVIAL	100% (6%)	75% (6%)	88% (14%)	94% (29%)	81% (28%)	88% (18%)
LOGICAL COLLISION-FREE	100% (4%)	65% (5%)	80% (12%)	83% (27%)	66% (27%)	74% (25%)
LOGICAL	99% (4%)	58% (4%)	78% (12%)	83% (26%)	67% (27%)	72% (25%)

Figure 5.5: Performance comparison of the CNN+GRU+FILM model on the different EXISTENTIAL caption patterns for various datasets, plus pattern distribution indicated in brackets. *"c"* refers to colour, *"s"* to shape, so *"c-only"* subsumes captions like *"A shape is red."*, whereas *"2s&c"* comprises statements like *"A shape is a square and a circle is red."*.

Visually colliding objects would explain why the shape attribute is more difficult to learn than the colour attribute, since the latter is not affected by overlap. However, the observed performance difference to COLLISION-FREE versions is small, and they exhibit the same tendency of colour mentions being beneficial whereas shape mentions detrimental. Furthermore, when comparing behaviour for agreeing versus disagreeing instances, no substantial differences can be recognised, so models do not struggle more with identifying disagreeing captions, or vice versa.

Comparing accuracy per shape in table 5.6 suggests that *"rectangles"*, and especially *"semi-circles"* and *"ellipses"* seem to be more difficult to learn correctly, by 2-4% and 4-6%, respect-

ively, based on EXISTENTIAL FULL (RELATIONAL-TRIVIAL and LOGICAL may contain multiple shape mentions per caption, so more difficult to analyse precisely). On the one hand, *"rectangles"* share contour features with *"squares"*, and *"semicircles"* as well as *"ellipses"* with *"circles"*, which could explain the differences. On the other hand, however, it is not clear why *"squares"* and *"circles"* are not affected equally. On the whole, performance per shape drops relatively uniformly from EXISTENTIAL ONE-SHAPE to LOGICAL.

	square	rectangle	triangle	pentagon	cross	circle	semicircle	ellipse
EXISTENTIAL ONE-SHAPE	100%	100%	100%	100%	100%	100%	100%	100%
EXISTENTIAL FULL	97%	95%	98%	98%	99%	98%	93%	93%
RELATIONAL-TRIVIAL	87%	82%	83%	86%	91%	88%	84%	84%
LOGICAL	73%	73%	71%	72%	72%	71%	70%	72%

Figure 5.6: Performance of the CNN+GRU+FILM model per shape type (all captions mentioning that shape) for different datasets.

Overall, the learning curves indicate that, aside from the general difficulty to process such statements correctly, all models find an equally effective sub-optimal heuristic quickly.

Models do not optimally generalise to the test distribution. Performance substantially decreases on the test data, which involves withheld shape-colour combinations. Interestingly, colour-only captions are still consistently solved, whereas performance of CNN+GRU+FILM for statements mentioning just the shape decreases to 86% on EXISTENTIAL ONE-SHAPE, 68% on EXISTENTIAL FULL and 54% on both RELATIONAL-TRIVIAL and LOGICAL, so basically chance level. Focusing only on statements with one of the withheld combinations, performance for all models is at 28-30%, which is worse than chance and thus indicates overfitting to training attribute combinations. Overall, the bias to judge a test instance as disagreeing is 62% for CNN+LSTM and 67-68% for the other models, whereas there is no such bias for the validation data.

In a follow-up investigation of the generalisation problem in table 5.7, four additional versions of the EXISTENTIAL COLLISION-FREE dataset (to avoid the negative influence of overlap) are generated, with an iteratively increasing set of test combinations by successively adding the multiple combinations of *"pentagons"* and *"grey shapes"*. Performance decreases from 84-85% to 79% after only adding the second pair of combinations, whereas validation accuracy stays at 99% throughout. The drop in test performance is mainly due to the lower accuracy for captions containing *"pentagon"* or *"grey"*, which are increasingly judged as disagreeing.

Detailed results for logical connectives. On the LOGICAL dataset, it is consistently observed that the best performance with 79-81% is achieved for statements with *"and"*, followed by

	six combinations	+ red pentagon + grey square	+ green pentagon + grey triangle	+ blue pentagon + grey circle	+ grey pentagon
Validation	99%	99%	99%	99%	99%
ratio agreeing vs disagreeing	0.97	1.00	1.00	1.03	0.97
Test	85%	84%	79%	79%	79%
ratio agreeing vs disagreeing	0.73	0.72	0.64	0.65	0.63

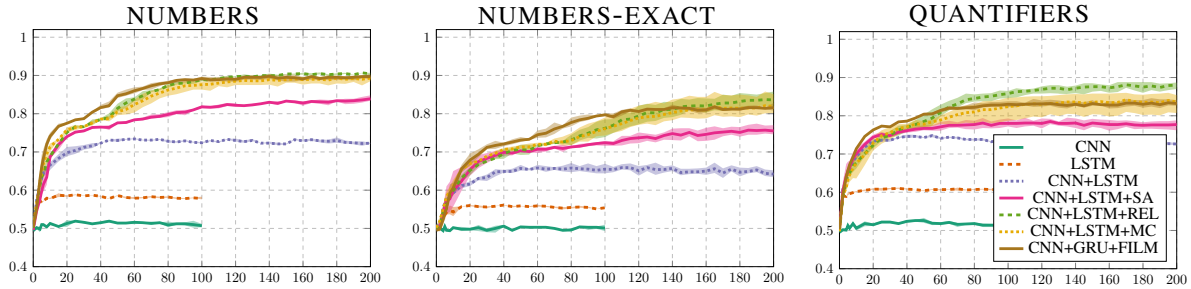
Figure 5.7: Performance of the CNN+GRU+FILM model on the validation set (following the same distribution as the training set) and test set, for an increasing number of withheld combinations (six/eight/ten/twelve/thirteen). Additionally, the ratio of agreeing versus disagreeing instances correctly classified as dis-/agreeing, indicating a bias towards incorrect on the test set.

73-77% for *"or"* and *"if"* (which, interestingly, can equivalently be defined via *"or"*). The connective *"if and only if"* seems to be the most difficult, with 64-65% accuracy, or 61% for CNN+LSTM.

Numbers, quantifiers and precise counting. Most non-baseline models reach an accuracy of 89-90% on the NUMBERS dataset after 200k iterations. Only the CNN+LSTM+SA model does not learn the dataset as well, with around 6% reduced accuracy. Performance of the CNN+LSTM baseline at 72% is worse still, but nonetheless substantially better than the unimodal baselines. Performance per number and modifier reveals further interesting patterns:

- *"zero"* instances are consistently the easiest to learn, followed by *"five"* (exception: CNN+GRU+FILM), while *"two"* is the most difficult.
- *"exactly"* instances are consistently the most difficult to learn, followed by *"no"*.
- Of the other modifiers, the ones including the bound in the threshold are consistently easier to learn, that is, *"at most"* is easier than *"less than"*, and *"at least"* is easier than *"more than"*.

The difficulty of statements involving *"exactly"* suggests that precise counting is more difficult than the form of approximate counting required for many of the NUMBERS instances. Indeed, accuracy levels decrease by 7-10% for the NUMBERS-EXACT variant, while relative differences between models do not substantially change.



In contrast, the differences between models are less pronounced for the QUANTIFIERS dataset: accuracy of CNN+LSTM is still at 73%, CNN+LSTM+SA with around 78% improves on that only by 5%, and the others by an additional 5-6%, or 10% in case of CNN+LSTM+REL. Of the quantifiers, *"all"* is easiest by a margin, followed by *"no"*, and *"a third"* or *"a quarter"* are the most difficult. Relative learning difficulties for modifiers are the same as mentioned above for NUMBERS.

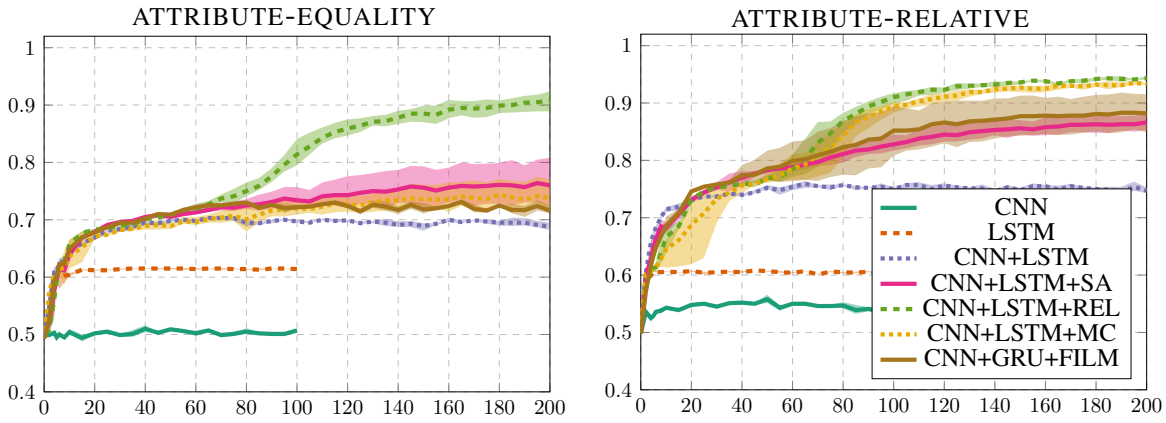
Table 5.8 visualises the order of difficulty across instance types for NUMBERS and QUANTIFIERS. While precise accuracy levels vary, the regularity in relative model performance is remarkable, since no model shows substantially different behaviour from the others in this perspective.

NUMBERS	lt	le	eq	ne	ge	gt	0	1	2	3	4	5
CNN+LSTM	78%	81%	64%	70%	77%	76%	80%	73%	69%	73%	75%	75%
...+SA	87%	89%	76%	81%	85%	85%	89%	83%	81%	82%	83%	85%
...+REL	93%	94%	86%	90%	92%	91%	94%	90%	90%	90%	91%	92%
...+MC	91%	92%	83%	88%	90%	89%	92%	87%	87%	88%	89%	91%
...+FILM	93%	94%	86%	89%	93%	91%	93%	91%	89%	90%	91%	91%

QUANTIFIERS	lt	le	eq	ne	ge	gt	no	1/4	1/3	1/2	2/3	3/4	all
CNN+LSTM	78%	82%	66%	71%	76%	75%	79%	71%	68%	73%	75%	80%	87%
...+SA	83%	85%	74%	74%	81%	81%	86%	75%	74%	79%	79%	83%	91%
...+REL	91%	93%	86%	86%	90%	88%	92%	86%	85%	89%	91%	91%	97%
...+MC	88%	91%	81%	83%	86%	86%	89%	82%	82%	85%	85%	89%	94%
...+FILM	87%	89%	80%	82%	86%	85%	89%	80%	81%	84%	85%	87%	93%

Figure 5.8: Performance comparison of the evaluated models on the different numbers/quantifiers and quantifier modifiers for the NUMBERS and QUANTIFIERS dataset. Colours indicate the relative order of performance per model on the different numbers/quantifiers/modifiers, from best to worst shifting from green towards red.

Non-spatial relational statements. The ATTRIBUTE-EQUALITY and ATTRIBUTE-RELATIVE datasets are both expected to require comparing the shape or colour attribute of two objects. Surprisingly, though, performance between the two datasets varies considerably. The CNN+LSTM+REL model performs best overall, and is the only model that substantially improves upon the others on the ATTRIBUTE-EQUALITY dataset, reaching an accuracy of 91%. The other non-baseline models here obtain only 72-76%, while the CNN+LSTM baseline is somewhat worse at 69% accuracy. For the ATTRIBUTE-RELATIVE dataset, CNN+LSTM+MC is on par with CNN+LSTM+REL with 93-94% accuracy, which both are 5-7% better than the other two non-baseline models. All of them substantially improve upon the 75% of the CNN+LSTM baseline.



An interesting observation is that some models achieve better results on these two datasets than on the supposedly simpler RELATIONAL-TRIVIAL dataset. Note, however, that in terms of shape/colour attributes these two datasets only require to keep track of either two shapes and one colour (*"same/different colour"* and *"darker/lighter"*), or one shape and two colours (*"same/different shape"* and *"smaller/bigger"*).

More detailed results for ATTRIBUTE-EQUALITY:

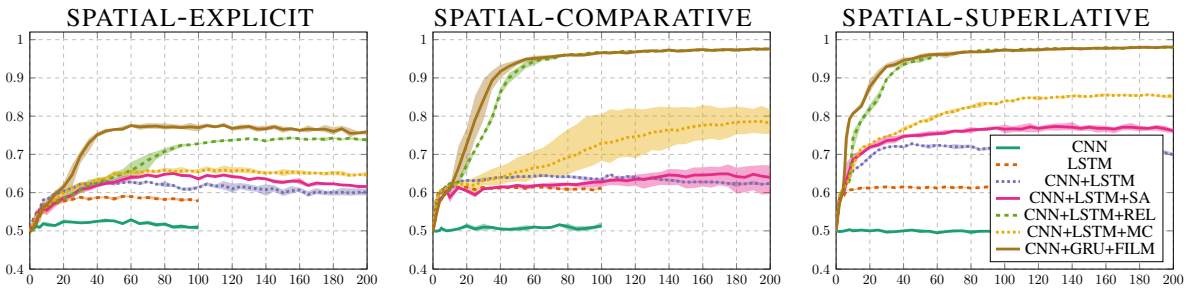
- Captions with *"different"* are handled with consistently higher accuracy than *"same"* statements, by a 5-13% margin for *"shape"* and 3-5% for *"colour"*.
- Except for the superior CNN+LSTM+REL model, statements containing *"different"* are solved with 6-9% higher accuracy for *"shape"* than for *"colour"*, while in the case of *"same"* the difference is 2-3% (exception: CNN+LSTM+SA).
- Performance for instances with *"different colour"* is consistently higher than for *"same shape"*, so the first effect (*"different"* being easier) is stronger than the second (*"shape"* being easier).

More detailed results for ATTRIBUTE-RELATIVE:

- Differences between overall performance are largely due to the instances with shade comparisons: *"smaller/bigger"* are processed with only 3-5% higher accuracy in the case of CNN+LSTM+REL/MC, while the gap increases to 9-15% for the other two non-baseline models, and to 21-23% for CNN+LSTM.
- There is no recognisable difference between the two comparisons for either shape (*"smaller/bigger"*) or colour (*"darker/lighter"*).

Spatial relations. Across all three SPATIAL datasets, the CNN+GRU+FILM model performs best, followed by CNN+LSTM+REL. On the SPATIAL-EXPLICIT dataset, CNN+GRU+FILM clearly dominates with an accuracy of 77% reached after only 50k iterations. CNN+LSTM+REL is the only other model which catches up later on and reaches a final performance of 74%, while

the others do not improve by more than 5% upon the CNN-LSTM baseline. On the two SPATIAL-IMPLICIT datasets, CNN+GRU+FILM and CNN+LSTM+REL show virtually the same learning curve and solve the dataset almost perfectly after around 60-80k iterations, with 97% final accuracy. The CNN+LSTM+MC model is second-best with an accuracy of 78% for SPATIAL-COMPARATIVE and 85% for SPATIAL-SUPERLATIVE, whereas CNN+LSTM+SA shows only slightly better performance than the CNN+LSTM baseline on SPATIAL-SUPERLATIVE.



Top performance is substantially lower on SPATIAL-EXPLICIT than on SPATIAL-IMPLICIT datasets. Considering that the relational inference required for both appears to be similar, this confirms that all models struggle with statements containing two distinct object descriptions.

More detailed results for SPATIAL-EXPLICIT:

- The superior performance of CNN+GRU+FILM and CNN+LSTM+REL mainly stems from improved learning of captions involving *"left/right/above/below"*, and to a lesser degree from *"closer/farther"* statements.
- With exception of the CNN+LSTM baseline, all models achieve a similar accuracy level of 70-75% for *"behind / in front of"* instances.
- All models are better in handling instances with agreeing captions, with accuracy differing by 7-11% compared to disagreeing captions.

More detailed results for SPATIAL-IMPLICIT:

- Focusing on CNN+LSTM+SA/MC, captions with *"closer/-est"* and *"farther/-est"* show consistently higher accuracy, by at least 3-4% on SPATIAL-SUPERLATIVE and by 13% on SPATIAL-COMPARATIVE.

A note on experiment runtimes. All experiments were run as a single-GPU job (Nvidia P100) on the Wilkes2 supercomputer of the High Performance Computing Service of the University of Cambridge. For the ShapeWorld experiments, most models took around 2-3 hours for 100k iterations. Only the CNN+LSTM+REL model took longer with 5-6 hours, due to the expensive computation across all pairwise combinations as part of its relation module. Compare this to the runtime for 100k iterations of the CLEVR experiments: 19 to 26 hours for most models when using raw images, but taking up to 35 to 40 hours in some cases, and ranging anywhere from 32 to around 50-55 hours and more when using pretrained image features.

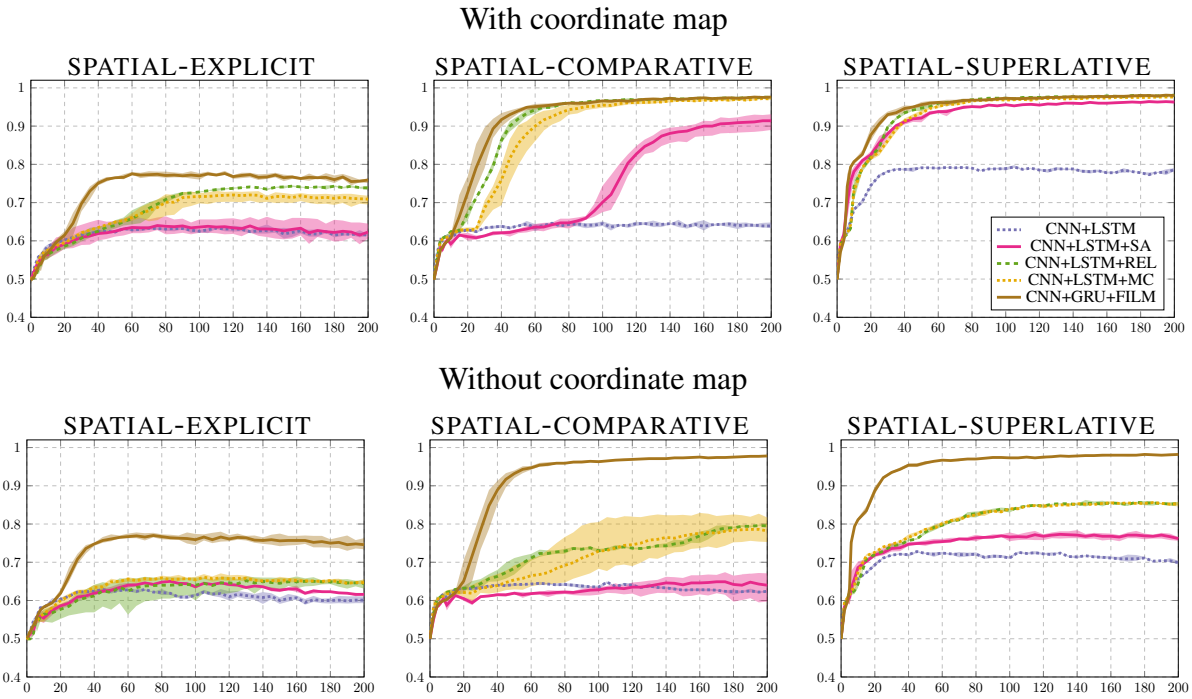
As a consequence of this considerable difference in runtime, not just are experiments faster and cheaper, but it enables a style of experimenting suitable to the unit-testing methodology: many small experiments per model with a quick response time that encourages iterative investigation (see also section 3.2.2 for a discussion of both aspects). Importantly, the experiments here demonstrate that CLEVR's already reduced complexity can be further simplified without trading off (in fact, arguably increasing) insights. This emphasises the value of starting with basic data and simple tests over more elaborate approaches with potentially unnecessary complexity.

5.5 Architecture analysis: priors for spatial reasoning[4]

One of the most surprising results of the experiments in section 5.4 are the substantial performance differences across models for the SPATIAL datasets. Since CLEVR contains questions with spatial relations like *"left of"* or *"in front of"*, my initial expectation was that the recent CLEVR models are able to correctly process phrases which require simple spatial inference. Another intriguing observation is that the two superior models, CNN+LSTM+REL and CNN+GRU+FILM, are also the ones where a map of relative spatial coordinates is attached to the visual features, which may be an advantage for processing spatial captions. Since this feature is unrelated to their core module, it raises the question: what architectural module(s) actually enable a model to handle spatial descriptions? In the following, I seek to shed light on this question for the three SPATIAL datasets. Simultaneously, this investigation is supposed to illustrate how, generally, unit-testing can inform model development by identifying what architectural modifications really contribute to improvements.

Spatial coordinate maps. Of the evaluated models, only two attach a map of relative spatial coordinates to the image features at the start of their respective core module: CNN+LSTM+REL and CNN+GRU+FILM. This minor detail infuses useful spatial information and thus relieves the core module of having to learn the concept of relative spatial positioning from scratch. The modification can easily be applied to the other models at the beginning of their respective core module as well, and the results below largely confirm the importance of this feature.

[4]Acknowledgements: The content of this section is also accepted and published as a paper entitled *"What is needed for simple spatial language capabilities in VQA?"*, co-authored with Ann Copestake, at the Visually Grounded Interaction and Language workshop of the Conference on Neural Information Processing Systems 2019 (Kuhnle and Copestake, 2019b).

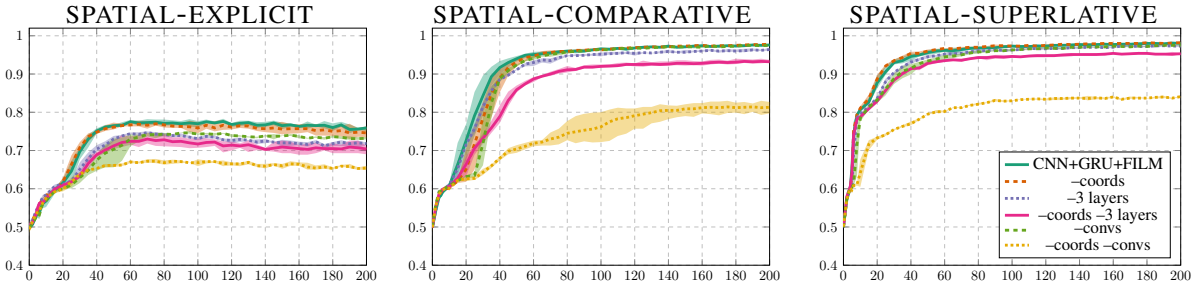


On the one hand, the modified CNN+LSTM+MC model is on par with CNN+LSTM+REL. CNN+LSTM+SA reaches a similar level to the other models only on SPATIAL-SUPERLATIVE, and a slightly worse level of 91% on SPATIAL-COMPARATIVE, while it does not improve on SPATIAL-EXPLICIT. Even the CNN+LSTM baseline profits from coordinates in the case of SPATIAL-SUPERLATIVE, improving by almost 10% despite the fact that positional image embeddings are pooled before being fused with language features. On the other hand, the accuracy of CNN+LSTM+REL without the spatial coordinates feature drops markedly. Only the performance of the CNN+GRU+FILM model remains virtually unchanged, implying that it does not (solely) rely on the coordinates to handle spatial data.

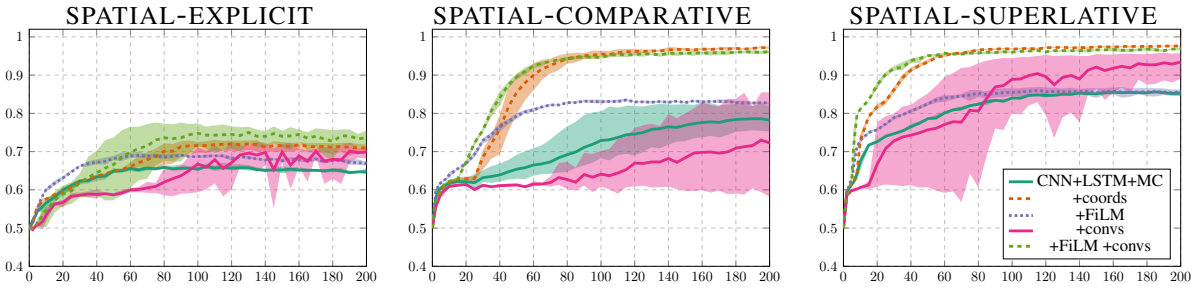
It can be concluded that spatial coordinates are a feature, though not the only one, which enables a model to handle SPATIAL instances. However, it requires early fusion of visual and language features and, of the evaluated core modules, it favours the simpler relation and multimodal core module over the more complex stacked attention layer to unfold its full potential. Note in particular that CNN+LSTM+REL and CNN+LSTM+MC show virtually the same behaviour in either condition. Since their architectures mainly differ in whether to process single positional image embeddings or pairwise concatenations thereof, the results indicate that the relation module – which is supposed to be an architectural prior beneficial for relational inference – does not contribute to improved performance here.

FiLM and convolutions. Another aspect which distinguishes the CNN+GRU+FILM core module from the other architectures is its use of convolutional layers with kernel size 3×3, instead of fully-connected layers applied independently per positional embedding. This allows the model to capture relative positions locally, and four subsequent such layers are enough

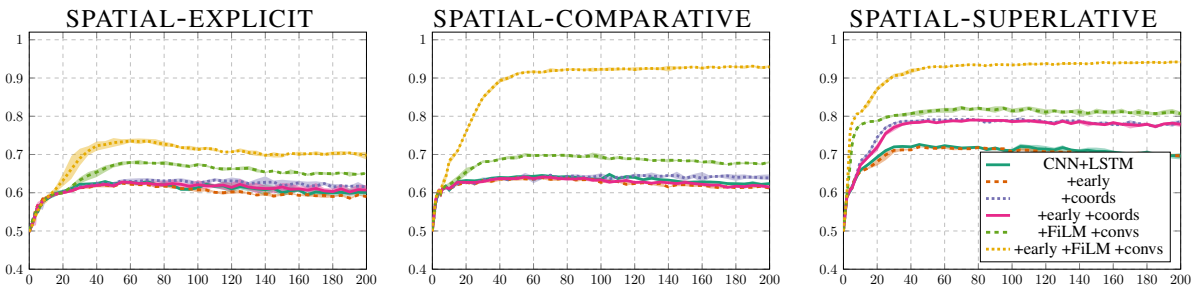
to cover the entire 8×8 feature space. Furthermore, CNN+GRU+FILM is the only model that fuses vision and language features four times – CNN+LSTM+SA does twice, all others once. However, as the ablation below shows, this feature, while beneficial, is not critical to its performance for spatial relations.



A first attempt to transfer this insight to the CNN+LSTM+MC model by simply replacing fully-connected layers with convolutions did not improve performance. It turns out that the beneficial effect of convolutions relies on feature-wise linear modulation to fuse language and visual features, as opposed to the otherwise typical concatenation. Results are shown below and confirm the effectiveness of the approach – in fact, CNN+LSTM+MC learns faster and achieves around 4% better performance on SPATIAL-EXPLICIT compared to the variant using coordinates.



These insights can even be transferred to the CNN+LSTM baseline. Since its late-fusion approach is not able to make full use of spatial information, results are also compared to an early-fusion variant, where the language embedding is combined with the visual features at the beginning of its core module via concatenation or feature-wise linear modulation. Surprisingly, early fusion does not at all affect performance of either the basic CNN+LSTM model or the variant with added coordinate map. However, while the combination of feature-wise linear modulation and convolutional layer already improves upon the other variants by 2-4%, combining it with early fusion boosts performance almost to the level of the CNN+GRU+FILM model: around 70% accuracy for SPATIAL-EXPLICIT, 93% for SPATIAL-COMPARATIVE and 94% on SPATIAL-SUPERLATIVE.



Note that replicating the feature is not possible for CNN+LSTM+REL, since the pairwise concatenation of positional embeddings destroys the two-dimensional arrangement of the image features, and consequently it is unclear how to apply convolutions here. Similarly, it is not clear how best to integrate both changes into the stacked attention layer of CNN+LSTM+SA.

Conclusion. Via feature ablation/addition and targeted architecture modification, two alternative techniques were identified whose presence or absence has a deciding impact on whether a model is able to achieve high performance on spatial ShapeWorld data: concatenating image features with relative spatial coordinates, or feature-wise linear modulation in combination with convolutions and early modality fusion. Other features, like the stacked attention layers of CNN+LSTM+SA or the relation module of CNN+LSTM+REL, did not on its own have a beneficial effect. Of the two alternative methods, adding a coordinate map to visual features is easier to integrate with any model architecture, whereas the combination of feature-wise linear modulation and convolutions is the more effective method.

5.6 Conclusion

The analysis of the selection of VQA models in this chapter has revealed substantial differences between models in some cases, besides identifying weak spots shared by all models. These weaknesses were not at all obvious from the literature, where most of the evaluated models have been shown to achieve roughly the same close-to-perfect level of performance on the CLEVR dataset, despite sharing the abstract domain of coloured shapes and diagnostic evaluation focus. Note that some of these findings may be specific to the choice of hyperparameters for each model and the data generator configurations used to produce the datasets. Although care has been taken to provide a thorough and fair comparison, further investigation could strengthen the claims, and confirm that the results are robust to changes in details of the experimental setup. However, this is ultimately an open-ended endeavour and as such beyond the scope of this chapter, whose main focus is to illustrate the proposed evaluation methodology and the ShapeWorld generation system.

The first two sections of chapter 6 present exploratory projects in which the original FiLM model CNN+GRU+FILM-ORIG is evaluated in more depth, and with differing results: model performance on relational statements is worse in section 6.1, and performance is close-to-perfect on quantifier statements in section 6.2. Some of the differences are likely due to the different hyperparameters and the fact that these experiments have been conducted earlier during my PhD, so with a slightly different dataset structure due to an older version of ShapeWorld. In particular section 6.1 confirms that differences in training dataset composition and distribution can have a huge impact on learnability of certain instance types.

On the one hand, the comparative assessment of a range of models in this chapter differs in focus from the in-depth evaluations of a single model in chapter 6, which is why the necessity for a unified hyperparameter setup emerged only here. On the other hand, the deviating results indicate a more fundamental problem: that hyperparameter tuning, even on clean abstract data, may not just tweak results slightly but can improve performance substantially – consider also the results in section 5.2 and 5.5 – and moreover that tuning the distribution of training data can have a similar impact. As a consequence, I want to emphasise that model performance results in this thesis are best understood comparatively, relative to other experiments in their context which follow the same setup, and not as absolute benchmark scores.

Other persistent findings are in accordance with the literature like, for instance, recent work suggesting that texture is more readily learned than shape (Geirhos et al., 2019), or that occlusion is problematic for object recognition (Rosenfeld et al., 2018). I thus recommend to use the texture as opposed to the colour attribute[5] and to turn off object collision for future experimentation based on ShapeWorld data, unless the investigation is supposed to focus on these weaknesses of current models.

I also want to emphasise that evaluation using a configurable data simulator like ShapeWorld is rarely perceived as 'complete', since the findings of one experiment frequently spark ideas for another set of experiments on slightly different data, encouraged by the fact that one just needs to adapt the generator configuration. The experiments in this chapter are supposed to provide a detailed comparative analysis on the breadth of ShapeWorld data, and thus only dig deeper into model behaviour in a few cases, while ignoring most other opportunities for further investigation. So whenever results have suggested obvious next steps, this should be seen as a strength of the evaluation approach and of a generic data simulator framework, that it both inspires ideas for more detailed analysis and provides the means to implement them.

[5] As mentioned in section 4.1, ShapeWorld supports texture attributes, but I ended up never using this feature, as this texture-preference 'problem' was not (widely) known until recently.

Chapter 6

Exploring other use cases for ShapeWorld

This chapter reports on three projects/papers which are influenced by my evaluation approach and based on the ShapeWorld data generation framework, but go beyond the comparative assessment of visual question answering models, as presented in chapter 5. Instead, these projects explore avenues for how the data and principles can be applied in another context or with a different evaluation focus. The following paragraphs briefly introduce each project. The last section 6.4 discusses other projects which have used ShapeWorld but on which I have not worked on myself. Besides hopefully inspiring similar investigations as future work, the purpose of this chapter is to illustrate the versatility and potential of a principled generation framework for evaluation data over ad hoc experiments and dataset creation.

The effect of multi-task and curriculum learning. Chapter 5 looked at different instance types independently, by training and evaluating models on a relatively targeted dataset in isolation. In the first project (section 6.1), I use ShapeWorld data to emulate a simplistic version of multi-task and curriculum learning. Both emphasise that a successful learning process may depend on the composition of the training data: multi-task learning tries to improve performance by simultaneously learning on a variety of 'tasks', and curriculum learning tries to bootstrap learning of difficult instances by iteratively increasing complexity over the course of training. My experiments are based on a 'narrow' interpretation of multi-task learning, according to which training on a broader dataset like CLEVR or the VQA Dataset, which comprise a variety of instance types and consequently subtasks, itself resembles a form of multi-task learning. However, it is not usually investigated to what degree the training process on such a monolithic dataset implicitly benefits from the concurrent presence of simpler 'pedagogical' instances. Experimental results suggest that learning behaviour is very sensitive to the compositional structure of the dataset and thus may have fundamental limitations.

Taking inspiration from psycholinguistics. The second avenue explored in this chapter (section 6.2) is the recently increasingly popular trend of borrowing experimental methodology

from cognitive psychology for deep learning models (Ritter et al., 2017; Nematzadeh et al., 2018). This approach acknowledges, on the one hand, that the capabilities of modern neural networks may be significantly more advanced than previous model classes and, on the other hand, that we have little understanding of their inner working and thus best treat them as black boxes. Facing similar conditions, psychologists study the behaviour of humans via carefully controlled experiments, usually not by observing random real-world situations, but by designing artificial environments and/or tasks that most clearly confirm or refute the investigated hypothesis. Methodologically, my version of unit-testing for deep learning shares many aspects with this approach. In this project, I show how a flexible data simulator like ShapeWorld makes it possible to replicate the psycholinguistic study of Pietroski et al. (2009) on how humans process the natural language quantifier *"most"*.

Towards better evaluation of image captioning. The third part of this chapter (section 6.3) investigates how the ShapeWorld framework can be leveraged to assess image captioning. Generative tasks are notoriously difficult to evaluate (as was already discussed in section 2.3) since, in comparison to discriminative tasks, the output space is vast and there is usually no well-defined expected response. However, ShapeWorld's grammar-based modelling of caption semantics makes it possible to verify precisely whether a caption formulates a true proposition given the content of the accompanying image. This capability enables accurate evaluation of image captioning models, which is effectively impossible to do with real-world datasets. In contrast, existing evaluation metrics for real-world datasets rely on a set of human-produced captions as a proxy for the content of an image. Limitations of the latter approach become particularly obvious in an abstract domain, thus confirming the value of the presented diagnostic approach as complementary to existing evaluation practice.

6.1 How clever is the FiLM model, and how clever can it be?

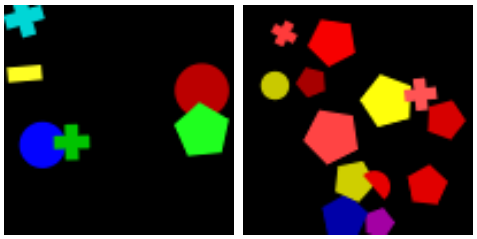
6.1.1 Introduction[1]

In this work, we analyse the learning process of the original FiLM model (Perez et al., 2018) in more detail. While FiLM manages to solve many tasks perfectly, it fails to achieve good performance on datasets involving relational statements in this investigation. We explore how two approaches – training on a broader 'multi-task' dataset including simpler instance types, as

[1] Acknowledgements: The results presented in this section are a continuation of the work Huiyuan Xie did as part of her MPhil thesis project in 2017/18, entitled *"How clever are the models exhibiting 'super-human' performance on the CLEVR VQA dataset?"*, which I proposed and co-supervised together with Ann Copestake as main supervisor. Nonetheless, the experiments and findings here represent my own work. The content of this section is also accepted and published as a paper of the same title, co-authored with Huiyuan Xie and Ann Copestake, at the Shortcomings in Vision and Language workshop of the European Conference on Computer Vision 2018 (Kuhnle et al., 2018). Since this section presents joint work, I will use plural forms like *"we"* instead of singular forms here.

EXISTENTIAL: *"There is a red square."*, *"A red shape is a square."*
EXISTENTIAL ONE-SHAPE: same as above, with only one object present
LOGICAL: two existential statements connected by: and, or, if, if and only if
NUMBERS: zero to five; with modifiers: less/more than, at most/least, exactly, not
QUANTIFIERS: with modifiers as above: no, half, all, a/two third(s), a/three quarter(s)
RELATIONAL: left, right, above, below, closer, farther, darker, lighter, smaller, bigger, same/different shape/colour (combination of ATTRIBUTE-EQUALITY, ATTRIBUTE-RELATIVE and SPATIAL-EXPLICIT)
SPATIAL TWO-SHAPES: the first four spatial relations, with only two objects per scene
RELATIONAL-NEGATION: relational plus negated relations
COMPARATIVE: left, right, upper, lower, smaller, bigger, darker, lighter, closer, farther (of two target objects)
SUPERLATIVE: superlative forms of the above, of an arbitrary number of target objects

Examples for visual scenes



Examples for true or false statements

- LOGICAL: *"There is a cyan square or a circle is green."*
- NUMBERS: *"At least two shapes are green."*
- QUANTIFIERS: *"More than half the pentagons are red."*
- RELATIONAL: *"A red cross is to the left of a yellow shape."*
- COMPARATIVE: *"The left circle is blue."*
- SUPERLATIVE: *"The lowermost yellow shape is a circle."*

Figure 6.1: *Top*: datasets together with their central words/constructions (see section 5.3 for more detail). *Bottom left*: two example visual scenes. *Bottom right*: example captions taken from different datasets, interpreted as 'questions' with corresponding yes/no answer depending on whether caption agrees with the image.

well as pretraining on simpler instances – can help alleviate these learning difficulties. However, we find that the multi-task approach is less robust than pretraining, and very sensitive to the compositional structure of the dataset.

These results put into question the common assumption of *"the effectiveness of data"* (Halevy et al., 2009) which underlies datasets such as the VQA Dataset (Antol et al., 2015), SQuAD for reading comprehension (Rajpurkar et al., 2016) or SNLI for language inference (Bowman et al., 2015): that all necessary abilities for a task can simply be learned from one big all-encompassing dataset, and that more data should lead to improved performance. Curriculum learning, on the other hand, shows promise as a robust approach to solving more complex instances of a task.

6.1.2 Experimental setup

Datasets. We generated various datasets based on existing ShapeWorld configurations. The different datasets are defined by the types of captions they contain. See figure 6.1 and section 5.3 for more details. Also note that the RELATIONAL dataset here comprises the two NON-SPATIAL datasets and SPATIAL-EXPLICIT. Each dataset consists of 500k training instances, plus 10k validation and test instances, respectively.

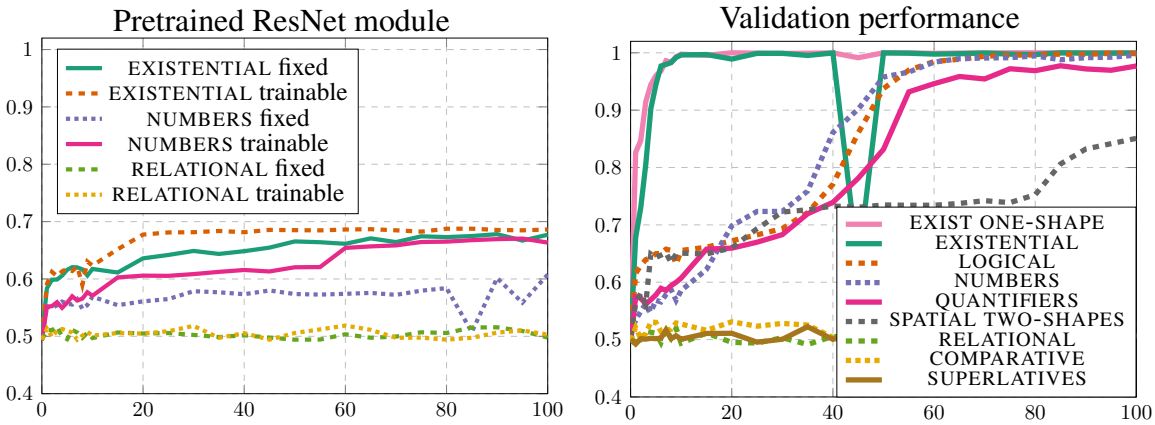


Figure 6.2: *Left diagram*: Performance when using a pretrained ResNet. *Right diagram*: validation performance of the FiLM model trained on various ShapeWorld datasets (*x-axis*: iterations in 1000, *y-axis*: accuracy).

Model. We focus on the FiLM model (Perez et al., 2018) here – more specifically, the original CNN+GRU+FILM-ORIG model (see section 5.1) with modified image module. The image is processed using a six-layer CNN (stride of 2 after the third and sixth layer) trained from scratch. We found that using a pretrained ResNet module did not perform well on our data. We train the model for 100k iterations in all experiments. Training performance is measured on the validation set every 1k iterations for the first 10k iterations and every 5k afterwards.

6.1.3 Results[2]

Pretrained ResNet does not perform well. We started off experimenting with the FiLM default of using a pretrained ResNet instead of a custom CNN. Versions with either a fixed or a trainable ResNet reach an accuracy of 65-70% after 100k iterations on EXISTENTIAL, which is substantially lower than the 100% when trained from scratch (see figure 6.2). This is surprisingly different from findings for CLEVR, where others reported the level of performance for either a pretrained ResNet or a custom CNN to be on a par (Perez et al., 2018; Santoro et al., 2017).

Many datasets solved and simple generalisation works. Overall, the FiLM model successfully learns many of our datasets (see figure 6.2). EXISTENTIAL is mastered after only 10k iterations and at the same speed as the trivial ONE-SHAPE variant. LOGICAL, NUMBERS and QUANTIFIERS reach close-to-perfect accuracy after around 60k iterations. The learning curves for these three tasks look remarkably alike and thus suggest a similar learning complexity for the model.

[2]See the discussion in section 5.6 about differences to results in chapter 5. Note in particular that the RELATIONAL here is a combination of three datasets from chapter 5, so results are not directly comparable.

Failure to learn relational statements. Surprisingly, we find that, with the exception of SPATIAL TWO-SHAPES, FiLM struggles to improve at all when trained on the various datasets requiring some form of relational reasoning (see figure 6.2): RELATIONAL, COMPARATIVE and SUPERLATIVE (referred to as RELATIONAL-LIKE below). The only exception is the simplistic TWO-SHAPES variant, but even here, learning is comparatively slow and, after plateauing for around 50k iterations at $\sim$75%, reaches only $\sim$85% after 100k iterations. This further emphasises the complexity for FiLM to learn relational statements.

Training on a broader set of instances. Datasets like CLEVR consist of a mix of instance types which require different understanding abilities, thus combining multiple tasks. Our assumption is that the simpler instances help to stabilise and guide the overall learning process, so that the more complex instances are also learned eventually[3], hence models are able to achieve close-to-perfect performance overall. We tested this assumption by training on broader combinations of datasets consisting of EXISTENTIAL, LOGICAL, NUMBERS, QUANTIFIERS plus some of the RELATIONAL-LIKE datasets (see figure 6.3). Indeed, FiLM is able to successfully learn multi-task datasets involving one of the more difficult datasets, or two in the case of COMPARATIVE and SUPERLATIVE. However, little to no improvement is observed in the other cases. These results further indicate that RELATIONAL seems to be the most complex of the RELATIONAL-LIKE datasets.

Augmenting with a simpler dataset. Additionally, we looked at the situation of a complex dataset paired with a simpler one, where instances of the latter can act as 'pedagogical' examples of a more general instance type. The FiLM model reaches $\sim$95% accuracy on a dataset augmenting the complex RELATIONAL with the simple TWO-SHAPES variant. However, performance stagnates when training on a combination with the more complex RELATIONAL-NEGATION instead of its negation-free variant (see figure 6.4).

Improvements by mixing/augmenting are unstable. Further investigation reveals that this 'synergy effect' of combining different datasets is very sensitive to the composition of the training set. On the one hand, FiLM fails to learn most multi-task datasets with two or more RELATIONAL-LIKE components as well as the augmented RELATIONAL-NEGATION dataset. On the other hand, even a slightly unbalanced distribution of 45% or 60% SPATIAL TWO-SHAPES with 55% or 40% RELATIONAL instances, respectively, shows no improvement above chance level (see figure 6.4).

The effectiveness of pretraining. In another series of experiments we investigated whether pretraining on simpler instances can bootstrap a successful learning process on more complex

[3]When referring to *"simple"* and *"complex"* or *"difficult"* instances here and in the following, we always mean with respect to the ability of the FiLM model to learn these instances.

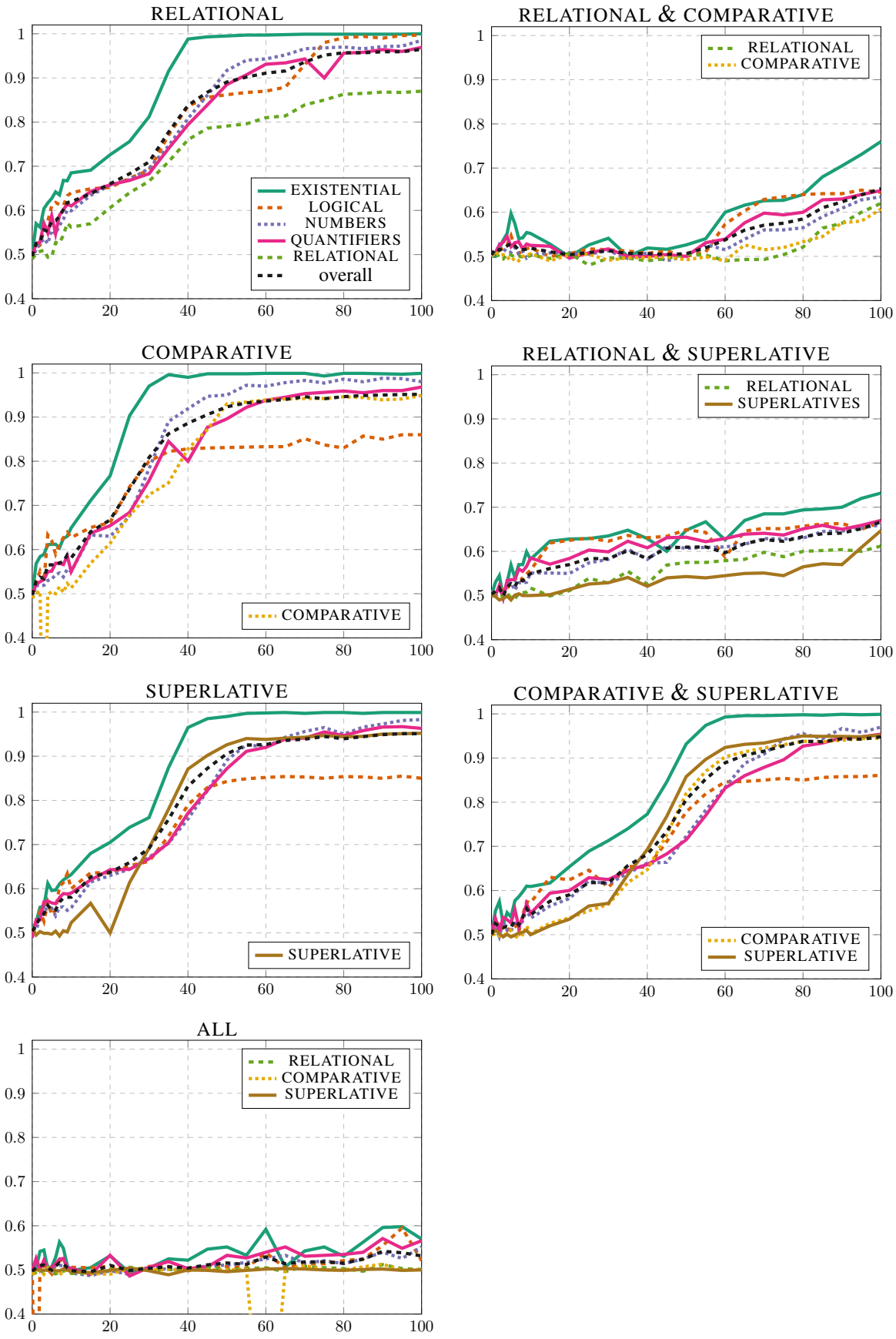


Figure 6.3: Performance per dataset of the FiLM model trained on broader combinations of datasets (*x-axis*: iterations in 1000, *y-axis*: accuracy).

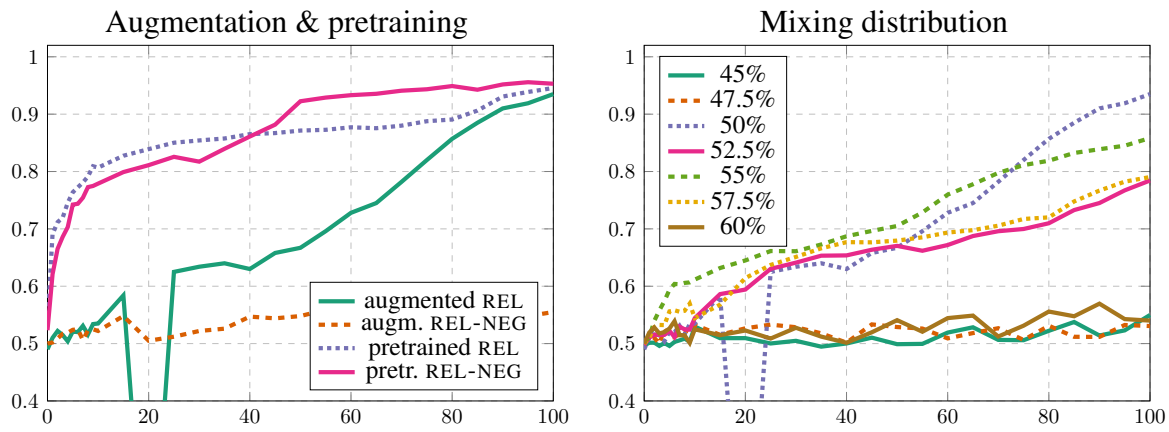


Figure 6.4: *Left diagram*: Performance on RELATIONAL/-NEGATION, when augmented with / pretrained on SPATIAL TWO-SHAPES. *Right diagram*: Distribution of SPATIAL TWO-SHAPES vs RELATIONAL instances (*x-axis*: iterations in 1000, *y-axis*: accuracy).

datasets, which is the assumption underlying curriculum learning (Elman, 1993; Bengio et al., 2009). For this, we take the model trained for 100k iterations on SPATIAL TWO-SHAPES and apply it to other RELATIONAL-LIKE datasets (see figure 6.4). For both RELATIONAL as well as RELATIONAL-NEGATION we observe a sharp increase in performance at the start, reaching $\sim$95% accuracy after 100k iterations. We particularly want to draw attention to the fact that the pretrained model reaches and eventually surpasses its previous performance level of $\sim$85% after only 20k/40k iterations, despite the more complex instances. Note also that the model trained on RELATIONAL-NEGATION seems to benefit from this dataset's increased complexity.

Differences to findings for CLEVR.

- Pretrained ResNet does not perform well.
- Simple compositional generalisation (simpler than CLEVR CoGenT) is learned perfectly.
- Relational statements are substantially more difficult to learn, at least in isolation.
- The presence of simpler instances likely benefits the learning of more complex ones.
- Performance on CLEVR does not transfer to all kinds of 'CLEVR-like' abstract data.

6.1.4 Discussion and conclusion

We have shown how the FiLM model struggles to learn relational statements when trained on a dataset of such statements only. Furthermore, we have investigated two mechanisms which help alleviate these difficulties: combining/augmenting training data with instances that are easier to learn, and pretraining on such simpler instances before moving to more complex ones. The first approach turns out to be very sensitive to the precise composition of the training set, while the second one leads to more robust improvements in our experiments.

In essence, combining various instance types ultimately results in big all-encompassing datasets for general tasks like VQA, where a variety of skills is assumed to be learned implicitly from a large number of input-output pairs. While our results confirm that this is possible (at least for synthetic data), they strongly question the robustness of this process. We showed how otherwise successful learning breaks down when the multi-task dataset is too complex or the mixing distribution is chosen wrongly. Note that these findings are based on clean and controlled abstract data, whereas the situation is even more complex for real-world datasets. Such sensitivity of the learning process to structural details of the training data is usually not considered, but might be able to explain some of the instability effects that are generally attributed to hyperparameter choice, random seeds, etc. Since it is hard to conceive how real-world data could ever be controlled to the degree possible with synthetic data, researchers should be more sceptical of complex architectures evaluated on only a single monolithic dataset, and instead encourage the reporting of negative results with respect to unstable performance and transfer failures.

Note that our findings are the result of a careful in-depth assessment of a single model for a range of instance types and configurations. We thus recommend to abandon the idea of 'datasets as tasks', and to shift focus from model building for an existing dataset to model analysis, by designing data and experiments which examine the learning behaviour in more detail.

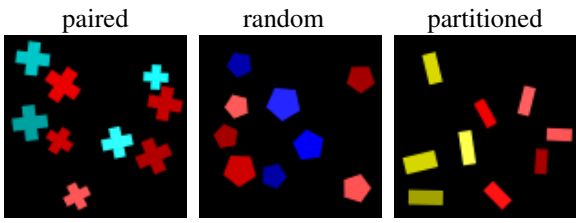
6.2 The meaning of "most" for visual question answering models

6.2.1 Introduction[4]

The correct interpretation of quantifier statements in the context of a visual scene requires non-trivial inference mechanisms. This work is inspired by experimental practice in psycholinguistics to shed light on the question how deep learning models for visual question answering learn to interpret statements involving the quantifier *"most"*. Following Pietroski et al. (2009), two strategies are discussed, which rely on fundamentally different cognitive concepts. By designing abstract visual scenes where the number and spatial arrangement of objects is controlled, it can be identified what strategy VQA models learn when trained on such data. Figure 6.5 illustrates how visual scenes can be configured to favour one over another mechanism. The experiments indicate that a form of approximate number system emerges, whose performance declines with more difficult scenes as predicted by Weber's law.

[4]Acknowledgements: The content of this section is also accepted and published as a paper of the same title, co-authored with Ann Copestake, at the BlackboxNLP 2019 workshop of the Annual Meeting of the Association for Computational Linguistics 2019 (Kuhnle and Copestake, 2019a).

I want to reiterate the difference of this experimental approach to mainstream machine learning practice (see also section 3.2). For different verification strategies, conditions are identified that should or should not affect their performance, and test instances are designed accordingly. By comparing the accuracy on various instance patterns, predictions about performance for these mechanisms can be verified and the most likely explanation identified. Note that the advocated evaluation methodology is entirely extrinsic and does not constrain the system in any way (like requiring attention maps) or require a specific framework (like being probabilistic).



"More than half the shapes are red shapes."

Figure 6.5: Three types of spatial arrangement of objects which may or may not affect the performance of a mechanism for verifying *"most"* statements. Going from left to right, a strategy based on pairing entities of each set and identifying the remainder presumably gets more difficult, while a strategy based on comparing set cardinalities does not.

Psychology as a discipline has focused entirely on questions around how humans process situations and arrive at decisions, and consequently has the potential to inspire a lot of experiments (like the ones presented here) for investigating the same questions in the context of machine learning. Similar to psychology, I advocate the preference of an artificial experimentation environment which can be controlled in detail, over the importance of data originating from the real world, to arrive at more convincing and thus meaningful results. However, for the replication of psychology experiments to deliver useful insights, they need to be properly 'translated' to the context of deep learning: in particular, one requires orders of magnitude more data, both for training and evaluation, as one cannot rely on the conformity of human common sense and world knowledge.

Taking inspiration from psychology seems particularly appropriate in the context of powerful deep learning models, which recently are frequently described by anthropomorphising words like *"understanding"* and compared to *"human-level"* performance. Instead of relying on the narrative of neural networks *"learning to understand/solve"* a task, researchers should corroborate their theories experimentally, that is, identify a reasoning mechanism that, if not human-like, at least is cognitively plausible. While this is by no means necessary for practically solving a task, I highlight two reasons why being able to comprehend model behaviour is nonetheless important: On the one hand, cognitive plausibility increases confidence in the abilities of a system – one is generally more willing to rely on a reasonable than an incomprehensible mechanism. On the other hand, pointing out systematic shortcomings inspires systematic improvements and hence can guide future progress. Moreover, particularly in the case of a human-centred domain like natural language, ultimately, some degree of comparability to human performance is indispensable.

6.2.2 Background: the meaning of "most"

Generalised quantifiers and "most". *"Most"* has a special status in linguistics due to the fact that it is the most prominent example of a quantifier whose semantics cannot be expressed in first-order logic, while other simple natural language quantifiers like *"some"*, *"every"* or *"no"* can directly be expressed in terms of the quantifier primitives $\exists$ and $\forall$ (plus logical operators $\wedge$, $\vee$ and $\neg$). Quantifiers like *"most"* require a fundamental extension of the logic system and its expressivity.

In the following, by x I denote an entity, A and B denote predicates (for instance, A as *"square"* and B as *"red"*), $A(x)$ is true if and only if x satisfies A (x is a square), and $S_A = \{x : A(x)\}$ is the corresponding set of entities satisfying this predicate (all squares). Thus the semantics of *"some"* and *"every"* can be defined:

$$\text{some}(A, B) \Leftrightarrow \exists x : A(x) \wedge B(x) \tag{6.1}$$

$$\text{every}(A, B) \Leftrightarrow \forall x : A(x) \Rightarrow B(x) \tag{6.2}$$

Importantly, these definitions do not involve the concept of set cardinality. This is not possible for *"most"*, which is commonly defined in one of the following ways:

$$\begin{aligned} \text{most}(A, B) &\Leftrightarrow |S_{A\wedge B}| > 1/2 \cdot |A| \\ &\Leftrightarrow |S_{A\wedge B}| > |S_{A\wedge\neg B}| \end{aligned} \tag{6.3}$$

"Most" is an example of a ***generalised quantifier***, and in fact all language quantifiers can be defined in terms of cardinalities, indicating the potential importance of a cardinality concept to human cognition.

Alternative characterisation. There is another way to define *"most"* which uses the fact that whether two sets are equinumerous can be determined without a concept of cardinality, based on the idea of a bijection:

$$A \leftrightarrow B :\Leftrightarrow \forall x : A(x) \Leftrightarrow B(x) \tag{6.4}$$

$$\Leftrightarrow |S_A| = |S_B| \tag{6.5}$$

The definition of equinumerosity can be generalised to *"more than"* (and correspondingly, *"less than"*), which lets us define *"most"* as follows:

$$\text{most}(A, B) \Leftrightarrow \exists\, S \subsetneq S_{A\wedge B} : S \leftrightarrow S_{A\wedge\neg B} \tag{6.6}$$

Although this definition has the same truth conditions as the one above, it suggests a different algorithmic approach to interpreting *"most"*, as I will discuss below.

Two interpretation strategies. The two characterisations of *"most"* are of course truth-conditionally equivalent, that is, every situation in which one of them holds, the other holds, and vice versa. Nevertheless, the subtle differences between these two characterisations suggest different algorithmic mechanisms of verifying or falsifying such statements, meaning that a system processes a visual scene differently to come to the (same) conclusion about a statement's truth.

Characterisation (6.3) represents the ***cardinality-based strategy*** of interpreting *"most"*:

1. Estimate the number of entities satisfying both predicates (*"red squares"*) and the number satisfying one predicate but not the other (*"non-red squares"*).
2. Compare these number estimates and check whether the former is greater than the latter.

I want to add that, actually, the two definitions in (6.3) already suggest a minor variation of this mechanism – see Hackl (2009) for a discussion on *"most"* versus *"more than half"*. However, I do not focus on this detail here, and assume the second variant in (6.3) to be 'strictly' simpler in the sense that both involve estimating and comparing cardinalities, but the first variant additionally involves the rather complex operation of halving one number estimate.

Characterisation (6.6) utilises the concept of a bijection, which corresponds to a comparatively simple pairing mechanism and as such could be imagined to be a primitive cognitive operation. This results in the ***pairing-based strategy*** of interpreting *"most"*:

1. Successively match entities satisfying both predicates (*"red squares"*) with entities satisfying one predicate but not the other (*"non-red squares"*).
2. The remaining entities are all of one type, so pick one and check whether it is of the first type (*"red square"*).

Cognitive implications. Finding evidence for one strategy over the other has substantial implications with respect to the 'cognitive abilities' of a neural network model. In particular, evidence for a cardinality-based processing of *"most"* suggests the existence of an **approximate number system** (ANS), which is able to simultaneously estimate the number of objects in two sets, and perform higher-level operations on the resulting number representations themselves, like a comparison operation. Explicit counting would be an even more accurate mechanism for this task, but neither is it available to the subjects in the experiments of Pietroski et al. (2009) due to very short scene display time, nor likely to be learned by the 'one-glance' feed-forward-style neural network evaluated in this work[5].

[5]By *"one-glance feed-forward-style networks"* I refer to the predominant type of network architecture which, by design, consists of a fixed sequence of computation steps before arriving at a decision. In particular, such models do not have the ability to interact with their input dynamically depending on the complexity of an instance, or perform more general recursive computations beyond the fixed recurrent modules built into their design. Important for the discussion here is the fact that precise – in contrast to approximate or subitising-style – counting is by definition a recursive ability, thus impossible to learn for such models.

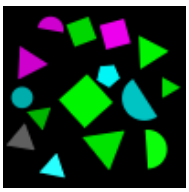
The ANS (see appendix in Lidz et al. (2011) for a summary) is an evolutionary comparatively old mechanism which is shared between many different species throughout the animal world. It emerges without explicit training and produces approximate representations of the number of objects of some type. They are approximate in the sense that their number judgement is not 'sharp', but resulting behaviour exhibits variance. This variance follows **Weber's law**, which states that the discriminability of two quantities is a function of their ratio[6]. The precision of the ANS is thus usually indicated by a characteristic value called **Weber fraction** which relates quantity and variance. The ANS of an adult human is reported to have an average Weber fraction of 1.14 or, more tangibly, it can distinguish a ratio of 7:8 with 75% accuracy. Finding evidence for the emergence of a similar system in deep neural networks indicates that these models can indeed learn more abstract concepts like approximate numbers than mere superficial pattern matching.

Both mechanisms for interpreting *"most"* suggest conditions in which they should perform well or badly. For the cardinality-based one, the difference in numbers of the two sets in question is expected to be essential: smaller differences, or greater numbers for the same absolute difference, require more accurate number estimates and hence make this comparison harder, according to Weber's law. The pairing-based mechanism, on the other hand, is likely affected by the spatial arrangement of the objects in question: if the objects are more clustered, pairing them with objects from the other set becomes harder. Importantly, these conditions are orthogonal, so each mechanism should not substantially be affected by the other condition, respectively. By constructing artificial scenes where one of the conditions dominates the configuration, and measuring the accuracy of being able to correctly interpret propositions involving *"most"*, the expected difficulties can be confirmed (or refuted) and thus indicate which mechanism is actually at work.

Using this methodology, Pietroski et al. (2009) show that humans exhibit a default strategy of interpreting *"most"*, at least when only given 200ms to look at the scene, and hence having to rely on an immediate subconscious judgement. This strategy is based on the approximate number system and the cardinality-based mechanism. Moreover, the behaviour is shown to be sub-optimal in some situations where humans would, in principle, be able to perform better, if they would deviate from their default strategy. Since machine learning models are trained by optimising parameters for the task at hand, it is far from obvious whether they learn a similarly stable default mechanism, or instead follow a (potentially superior) adaptive strategy depending on the situation. The former would suggest that the system is able to acquire and utilise core concepts like an approximate number system.

[6]There is evidence for Weber's Law in a range of other approximate systems, some of them non-discrete and thus rendering a pairing-based strategy impossible. While this does not rule out such a strategy when observing performance decline as predicted by Weber's Law, it strongly suggests that similar and thus non-pairing-based mechanisms are at work in all of these situations.

- Exactly two squares are yellow.
- Exactly no square is red.
- More than half the red shapes are squares.
- More than a third of the shapes are cyan.



- Less than half the shapes are green.
- Exactly all magenta shapes are squares.
- At most five shapes are magenta.
- At least one triangle is grey.

Figure 6.6: Two example images with four in-/correct captions each from the Q-FULL dataset.

One may speculate about the innate preference of modern network architectures for either of the strategies: Most of the visual processing is based on convolutions which, being an inherently local computation, I assume would favour the pairing-based strategy via locally matching and 'cancelling out' entities of the two predicates. On the other hand, the tensors resulting from the sequence of convolution operations are globally fused into a final embedding vector, which in turn would support the more globally aggregating cardinality-based strategy. However, the type of computations and representations learned by deep neural networks are poorly understood, making such speculations fallacious. I thus emphasise that the higher-level motivation for this work is to demonstrate how one needs not rely on such 'speculations', but can experimentally substantiate such claims.

6.2.3 Experimental setup

The setup closely resembles the psychological experiments conducted by Pietroski et al. (2009), but aimed at a state-of-the-art VQA model.

I experiment with two different training datasets: Q-FULL is based on both QUANTIFICATION captioners implemented in ShapeWorld, NUMBERS and QUANTIFIERS (see section 5.3), whereas Q-HALF is restricted to only the two quantifiers *"more than half"* and *"less than half"*. Figure 6.6 shows two images together with potential Q-FULL captions.

However, the existing world generator modules are too generic for my evaluation purposes here, since they do not allow to control attributes and positioning of objects to the desired degree. Consequently, I implement a new custom generator module with the following functionality to produce test data.

Attribute contrast: For each instance, either the attribute *"shape"* or *"colour"* is picked[7], and subsequently two values for this attribute and one value for the other are randomly chosen. This means that the only relevant difference between objects in every image is either the shape or colour value (for instance, *"squares vs circles"* or *"red vs blue shapes"*).

Contrast ratios: A list of valid ratios between the contrasted attributes can be specified, from which one will be randomly chosen per instance. For instance, a ratio of 2:3 means that there are 50% more objects with the second than the first attribute. The values of interest

[7]Note that the examples in figures were chosen to always vary in colour, so differences are more easily visible.

Figure 6.7: From left to right, the ratio between the two attributes is increasingly balanced.

are close to 1:1, that is, 1:2, 2:3, 3:4, 4:5, etc. The increasing difficulty resulting from closer ratios is illustrated in figure 6.7. Multiples of the smaller ratios are also generated (e.g., 2:4 or 6:9 in case of 1:2), within the limit of up to 15 objects overall.

Area-controlled (vs size-controlled): If this option is set, object sizes are not chosen uniformly across the entire valid range, but size ranges are chosen so that both attributes cover the same image area on average. This means that the more numerous attribute will generally be represented by smaller objects and, moreover, that the difference in covered area between, for instance, squares and triangles is taken into account.

While objects are positioned randomly by default, two generator modes are implemented which control this aspect as well. Figure 6.5 in the introduction illustrates the different modes.

Partitioned positioning: An angle is randomly chosen for each image, and objects of the contrasting attributes are consistently placed either on one or the other side.

Paired positioning: If there are objects of the contrasted attribute which are not yet paired, one of them is randomly chosen and the new object is placed next to it.

The captions of these evaluation instances are always of the form *"More/less than half the shapes are X"."* with *"X"* being the attribute in question, for instance, *"squares"* or *"red"*. Note that this is an even more constrained captioner than the one used for Q-HALF, since the subject is always fully underspecified as *"shape"*. I also emphasise that, in contrast to these targeted test configurations, the default generator is used to generate the training instances in Q-HALF and Q-FULL. So these images generally contain many more than just two contrasted attributes, and ratios between attributes are not controlled for. The examples in figure 6.6 are chosen to illustrate this fact: the second example contains a *"half"* statement with ratio 7:8, whereas the first contains one about a 0:4 ratio even though the image would also allow for a more 'interesting' 3:4 ratio (colour of semicircles).

While I overall try to stay close to the experimental setup of Pietroski et al. (2009), in the following I point out some differences. Most importantly, instead of just using yellow and blue dots, I use all eight shapes and seven colours that ShapeWorld provides. This increases the visual variety of the instances, and thus encourages the system to actually learn the fact that shape and colour are attributes that can be combined in any way. Note that humans in psychological experiments have learned language in even more complex situations, which

cannot be approximated here. Moreover, the data does not contain yes/no *"most"*-questions, but true/false captions with equivalent phrasings *"more/less than half"*. Since the model is trained from scratch on such data, this should not affect results.

I do not implement their *"column pairs mixed/sorted"* modes since they would result in comparatively big and mostly empty images, hence require bigger networks and might cause practical learning problems due to sparseness, which are not supposed to be addressed here. In contrast, the partitioned mode is more difficult than the ones investigated by Pietroski et al. (2009), at least for a pairing-based mechanism.

Model. I focus on the FiLM model (Perez et al., 2018) here – more specifically, the original CNN+GRU+FILM-ORIG model (see section 5.1) with modified image module. The image is processed using either the pretrained ResNet-101, or a variant of the version trained from scratch on raw images, which consists four convolutional layers with the second and fourth using a stride of 2.

Training details. The training set for both Q-FULL and Q-HALF consists of around 100k (25x 4096) images with 5 captions per image, so overall around 500k instances. The model is trained for 100k iterations with a batch size of 64. Training performance is measured on an additional validation set of 20k instances. Moreover, 1024 instances for each of the overall 48 evaluation configurations are produced, to investigate the trained model in more detail.

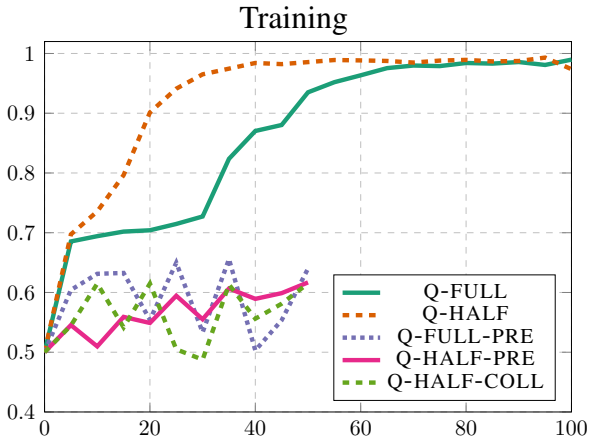


Figure 6.8: Training performance (*x-axis*: iterations in 1000, *y-axis*: accuracy). Q-FULL: unconstrained dataset; Q-HALF: dataset restricted to *"less/more than half"*; -PRE: using pretrained CNN module; -COLL: allowing object overlap.

6.2.4 Results[8]

Training. Two versions of the FiLM model are assessed, with the CNN module trained from scratch on the task: one on the Q-FULL dataset which contains all available quantifier and number caption types, the other on the Q-HALF dataset which is restricted to only captions involving the quantifier *"half"*. Performance of the system over the course of the 100k training iterations is shown in figure 6.8. The two models, referred to by Q-FULL and Q-HALF below, learn to solve the task quasi-perfectly, with a final accuracy of 98.9% and 99.4%

[8] See the discussion in section 5.6 about differences to results in chapter 5.

respectively. Not surprisingly, the system trained on the more diverse Q-FULL training set takes longer to reach this level of performance, but nevertheless plateaus after around 70k iterations.

For the sake of completeness, the performance of other models which failed to show clear improvement over the first 50k iterations is also included in this figure. This includes the model with pretrained instead of trainable CNN module (Q-FULL-PRE, Q-HALF-PRE), and an earlier trial on Q-HALF where the data generation was not constrained to not produce object collisions (Q-HALF-COLL, the default in ShapeWorld is to allow up to 25% area overlap).

Evaluation. Table 6.9 presents a detailed breakdown of system performance on the evaluation configurations. Before discussing the results in detail, I want to reiterate three key differences between the evaluation data and the training data:

- The visual scenes all exhibit close-to-balanced contrast ratios, while this is not the case for the training instances.
- The evaluation scenes only contain objects of two different attribute pairs, and consequently the numbers to compare are generally greater than in the training instances, where more attributes are likely present in a scene.
- Q-FULL contains more than just statements involving *"half"* – in fact, a random sample of 100 images and 500 captions suggests that they constitute only around 8% of the dataset (and this includes combinations with modifiers beyond *"more/less than"*).

Considering these differences, the relatively high accuracy on test instances throughout indicates a remarkable degree of generalisation.

More balanced ratios. The most consistent effect is that more balanced ratios of contrasted attributes cause performance to decrease. This is certainly affected by the tendency of the training data to not include many examples of almost balanced ratios. However, if this were the only reason, one would expect a much more sudden and less uniform decrease. More importantly, since Q-FULL generally contains fewer *"half"* statements, the decline should be more pronounced here. Neither of these effects is observed, and it can hence be concluded that both models have actually developed a more sophisticated mechanism than superficial pattern matching. This is further discussed at the end of this section.

Random vs paired vs partitioned. There is a clear negative effect of the partitioned configuration on performance for the model trained on Q-FULL, which suggests that the learned mechanism is not robust to a high degree of per-attribute clustering. This indicates at most a weak preference towards a pairing-based strategy for Q-FULL, though, since otherwise the model would not be expected to perform best on the random configuration. Interestingly, the results for Q-HALF even suggest slightly better performance on the area-controlled partitioned

train	mode	size-controlled								area-controlled							
		all	1:2	2:3	3:4	4:5	5:6	6:7	7:8	all	1:2	2:3	3:4	4:5	5:6	6:7	7:8
Q-FULL	random	92	100	99	97	94	91	88	85	93	100	99	97	93	91	86	82
	paired	93	99	99	96	93	90	88	82	93	99	99	96	91	87	84	80
	part.	89	100	99	92	90	81	77	72	89	99	98	92	88	82	78	72
Q-HALF	random	92	100	100	98	93	88	88	87	93	100	100	97	92	86	85	82
	paired	92	100	100	96	90	86	84	79	92	100	99	96	87	84	79	76
	part.	91	100	99	96	86	83	83	80	91	100	99	94	89	83	83	80

Figure 6.9: Accuracy of the model when trained on either Q-FULL or Q-HALF for the various evaluation configurations.

configuration. Overall, no clear preference for either the perfectly clustered partitioned or the perfectly mixed paired arrangement is apparent. Note, however, that the random mode instances are most similar to the random placement of objects in the training data, which might cause this preference.

Size- vs area-controlled. The performance in both cases is comparable, showing that the models do not solely learn to rely on comparing the overall covered area, which would only work well in the size-controlled mode. Nevertheless, a tendency is observed for area-controlled instances to be somewhat more difficult in random and paired mode, more so for Q-HALF, which suggests that the models may learn to use covered area as a feature to inform their decision in some cases.

Q-FULL VS Q-HALF. There seems to be a tendency of the system trained on Q-FULL to perform marginally better, except for the partitioned mode discussed before. The fact that this model performs at least on a par with the one trained on Q-HALF, while only seeing a fraction of directly relevant training captions, indicates that the learning process is not 'distracted' by the variety of training captions, and indeed might profit from it.

Ratios and Weber fraction. I generated evaluation sets of even more balanced ratios (8:9, 9:10, 10:11, increasing the overall number of objects accordingly to 17/19/21), and in figure 6.10 plotted the accuracy of the Q-FULL model on increasingly balanced sets for all three spatial configuration modes, not controlling for area (which for greater numbers only has a negligible effect anyway). The figure also contains a diagram with accuracy plotted against ratio fraction, which is more common in the context of Weber's law. The characteristic Weber fraction can be read off directly as the ratio at which a model is able to distinguish two values with 75% accuracy: around 1.11 for random/paired and 1.16 for partitioned, which corresponds to 9:10 and 6:7 as closest integer ratios. These values are in the same region as the average human Weber fraction, which is often reported as being 1.14, or 7:8.

I emphasise that these curves align well with the trend predicted by Weber's law, even for the ratios with more than 15 objects overall, where such situations have never been encountered

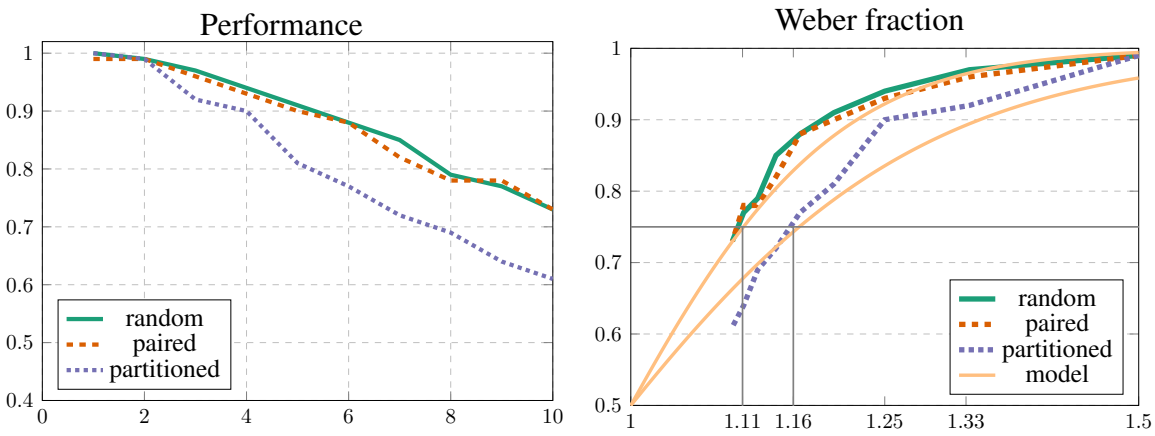


Figure 6.10: *Left*: Q-FULL model performance for increasingly balanced ratios (x-axis: n for ratio n:n+1). *Right*: Performance as a function of the actual ratio n+1/n, with Weber fractions (75%) highlighted, plus corresponding idealised model Weber curves.

during training. All this strongly suggests that the model learns a mechanism similar to an approximate number system, which is able to produce representations that can be utilised for identifying the more numerous set. In particular, it can be concluded that the system does not actually learn to explicitly count, since one would then not expect to observe such fuzziness characteristic to an approximate system.

Moreover, since performance is affected somewhat by the partitioned and the area-controlled modes, the interpretation of *"most"* seems to be informed by other features as well. As noted earlier, since the model is trained to optimise this task, development of an adaptive strategy is not unexpected. On the contrary, more surprising is the fact that an ANS-like system seems to emerge as a primary 'backbone' mechanism, with additional factors acting as less influential secondary features.

6.2.5 Related work on numbers, quantifiers and counting

The VQA Dataset (Antol et al., 2015) provides a shallow categorisation of questions, including basic count questions, however, these categories are far too coarse for a similar investigation as presented here. CLEVR (Johnson et al., 2017a) covers some abilities like numbers or attribute comparisons in more detail, but still in a fixed categorisation. More recently, the COG dataset (Yang et al., 2018) was introduced, which most explicitly focuses on replicating psychological experiments for deep learning models, hence most related to this work. However, their dataset does not contain any number or quantifier statements.

There is some work on investigating deep neural networks which look at numerosity from a more psychologically inspired viewpoint. Stoianov and Zorzi (2012) found that visual numerosity emerges from unsupervised learning on abstract image data. Zhang et al. (2017b) looked at salient object subitising in real-world images, formulated as a classification task over five classes ranging from *"0"* to *"4 or more"*. In a more general number-per-category classification

setup, Chattopadhyay et al. (2017) investigated different methods of obtaining counts per object category, including one which is inspired by subitising. Moving beyond explicit number classification, (Zhang et al., 2018c) recently introduced a dedicated counting module for visual question answering.

Another line of work looked at a similar classification task, but for proper quantifiers like *"no"*, *"few"*, *"most"*, *"all"*, first on abstract images of circles (Sorodoc et al., 2016), then on natural scenes (Sorodoc et al., 2018). Recently, Pezzelle et al. (2018) investigated a hierarchy of quantifier-related classification abilities, from comparatives via quantifiers like the ones above to fine-grained proportions. Wu et al. (2018), besides investigating precise numerosity via number classification as above, also look at approximate numerosity as binary greater/smaller decision, which closely corresponds to the experiments here. However, on the one hand, their focus is on the subitising ability, not the approximate number system. On the other hand, their experiments follow a different methodology in that they already train models on specifically designed datasets, while I deliberately leverage such targeted data only for evaluation.

On a methodological level, my proposal of inspiring experimental setup and evaluation practice for deep learning by cognitive psychology is in line with that of Ritter et al. (2017) and their shape bias investigation for modern vision architectures (see also section 2.4.2).

A few months after our paper was first published on arXiv (Kuhnle and Copestake, 2019a), O'Sullivan and Steinert-Threlkeld (2019) reported on very similar work, where they also replicated the study of Pietroski et al. (2009) to analyse how deep learning models process the quantifier *"most"*. However, their investigation differs in two ways: on the one hand, their focus is on the potential of using neural networks as cognitive models for such tasks and, on the other hand, they explicitly operationalise the concept of task duration.

6.2.6 Conclusion

Two strategies of algorithmically interpreting *"most"* in a visual context are identified, with different implications on cognitive concepts. Following experimental practice of similar investigations with humans in psycholinguistics, experiments and data are designed to shed light on the question whether a state-of-the-art VQA model shows preference for one strategy over the other. Performance on various specifically designed instances does indeed indicate that a form of approximate number system is learned, which generalises to more difficult scenes as predicted by Weber's law. The results further suggest that additional features influence the interpretation process, which are affected by the spatial arrangement and relative size of objects in a scene. There are many opportunities for future work from here, from strengthening the finding of an approximate number system and further analysing confounding factors, to investigating the relation to more explicit counting tasks, to extending the evaluation to other visual question answering models which also exhibit good performance on related tasks (Hudson and Manning, 2018; Zhang et al., 2018c; Santoro et al., 2017).

6.3 Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity

6.3.1 Introduction and motivation[9]

Image captioning as a multimodal task has drawn much interest in recent years. However, evaluation for this task – as well as other generative tasks (more details in section 2.3) – remains a challenging problem. The main difficulty is that generative tasks are by their nature less constrained with respect to the expected response, in contrast to discriminative tasks where data points can usually be consistently annotated with ground-truth gold labels.

Researchers have nonetheless proposed automatic evaluation metrics based on annotated datasets, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE (Lin, 2004), CIDEr (Vedantam et al., 2015) or SPICE (Anderson et al., 2016). These metrics use a set of human-produced example captions as approximate representation for what 'correct' captions are supposed to look like, and compare similarity of the output of a captioning system to this reference, often based on n-gram overlap, or propositional triples extracted from a semantic graph parse in the case of SPICE. Consequently, they do not check the actual relation between a candidate caption and the visual target, but take a set of 'ground-truth' statements as proxy for image content.

Caption 1: A circle is above a green rectangle.

Caption 2: A blue triangle is to the left of a semicircle.

Caption 3: A semicircle is below a grey triangle.

Caption 4: A semicircle is to the left of a triangle.

Figure 6.11: ShapeWorld example: spatial statements in the context of multiple shapes. The first three statements are truthful and diverse descriptions of the image. The fourth statement is wrong, but nonetheless exhibits a high degree of n-gram overlap with the true reference captions.

For real-world datasets, the set of reference captions is often relatively coherent, and thus gives the illusion that they indeed approximate a more or less well-defined ideal caption space – unsurprisingly, since many datasets consist of photographs whose composition is purposefully

[9]Acknowledgements: The work on ShapeWorld for image captioning started with the MPhil thesis project of Tom Sherborne in 2017/18, entitled *"Evaluating image description systems with truth-conditional semantics"*, which I proposed and co-supervised together with Ann Copestake as main supervisor. Subsequently, the work was continued mainly by Huiyuan Xie during the first year of her PhD in 2018/19, in collaboration with Ann Copestake and I, which resulted in a paper with the same title as this section by Huiyuan Xie, Tom Sherborne, Ann Copestake and I, accepted and published at the Evaluating Evaluation of AI Systems workshop of the AAAI Conference on Artificial Intelligence 2020 (Xie et al., 2020). Since the experimental work was done by Huiyuan Xie and Tom Sherborne, I will only briefly report the results here, and instead focus on my contribution to the project, which are mostly related to the motivation behind the project and the parts related to ShapeWorld. Moreover, since this section presents joint work, I will use plural forms like *"we"* instead of singular forms here.

chosen. However, moving to the abstract domain of ShapeWorld, where no object or relation is naturally more noteworthy than others, it becomes apparent that this methodology works only under specific conditions. Consider the example ShapeWorld instance in figure 6.11: the first three captions are true statements about the image and express relevant ideas, but describe different objects, attributes and relationships, whereas the fourth caption is wrong despite referring to the same objects as the third caption. While existing metrics have undeniably been useful for image captioning evaluation, the example illustrates that their focus on surface similarity limits their ability to provide deeper insights into learned model behaviour.

As I have already discussed in section 4.5, thanks to the MRS formalism and the ERG grammar being bidirectional, the ShapeWorld system implementation is not restricted to generating captions, but can easily be extended to instead parse and analyse examples. The goal of this project is to extend the scope of the ShapeWorld's diagnostic evaluation framework to the generative task of image captioning, with the goal of directly assessing model output for grammaticality, truthfulness and diversity. Besides the parsing functionality, most components of the framework can be reused: the generation of training data can simply be configured accordingly, in particular correct-only captions, and the assessment of whether a parsed caption model applies to an image is already implemented as part of the generation process (see section 4.2).

In contrast to discriminative tasks where the difficulty lies in generating interesting problem instances, for which model performance would clearly indicate strength or weakness, image captioning as a generative task instead requires the ability to precisely assess correctness and distinguish appropriateness in the broad space of possible model outputs. A grammar-based data generator like ShapeWorld provides the means to do both. Note, however, that image captioning evaluation as proposed here is only possible if data includes a model representation with sufficient information to enable a complete and correct analysis of the model output – or at least an approximation thereof, as in Madhyastha et al. (2019). As with visual question answering, unit-testing for image captioning is proposed as a complementary evaluation step in addition to assessing model behaviour on real-world data.

6.3.2 GTD evaluation framework

We propose a set of principled evaluation criteria which evaluate image captioning models for grammaticality, truthfulness and diversity (GTD). These criteria arguably correspond to necessary requirements for image captioning systems: (a) that the output is grammatical, (b) that the output statement is true with respect to the image, and (c) that outputs are diverse and mirror the variability of training captions.

Grammaticality. An essential criterion for an image captioning model is that the generated captions are grammatically well-formed. Assessing grammaticality in a general context is itself a difficult task, but becomes more feasible in a very constrained context like our diagnostic

language data. We take parseability with the English Resource Grammar (ERG, see section 4.3 for more information) as a surrogate for grammaticality, meaning that a sentence is considered 'valid' if it is possible to reverse-engineer a corresponding ShapeWorld caption representation, and the associated metric is the ratio/accuracy of valid output sentences. Considering the highly regular setting of ShapeWorld and the fact that training data – the only language source for the model to learn from – is generated using the same grammar, the ERG has $\sim$100% coverage in the model output space. In contrast, metrics like BLEU implicitly approximate grammaticality as n-gram overlap with a set of reference captions, which conflates grammaticality and semantic assessment in one final performance score.

Truthfulness. The second aspect we investigate is truthfulness, that is, whether a candidate caption is compatible with the content of the image it is supposed to describe. In the ShapeWorld framework, before realising caption objects as natural language via the ERG, their logical semantics is evaluated against the abstract world model, to ensure that reference captions are true descriptions of the corresponding visual scenes. We use this capability of the system to test whether the grammatical captions – the ones which can be parsed – agree with the visual content of the image, and the associated metric is the ratio/accuracy of agreeing output captions. Truthfulness is consequently the relative ShapeWorld semantics is equivalent to literal and context-agnostic language interpretation in the style of traditional formal semantics, which is adequate for assessing the truthfulness of captions. In comparison to other image captioning metrics, we do not rely on a set of captions as a surrogate for the content of an image, but instead leverage the fact that the ground truth is available, thus enabling the evaluation of actual image-caption agreement.

Diversity. While grammaticality and truthfulness are essential requirements for image captions, these criteria alone can easily be 'gamed' by specialising on a small set of generic statements which are true most of the time. In the context of abstract shapes, such captions include examples like *"There is a shape."* or *"At most five shapes are blue."* (which is technically true even if there is no blue shape). This motivates the third fundamental requirement of captioning output to be diverse. Since ShapeWorld produces caption content randomly, we take the reference captions accompanying the test images as a proxy for optimal diversity, and compare it with the empirical output diversity of the evaluated model on these test images. Practically, we look at the number of distinct language constructions, and compute the diversity score as the ratio of observed versus the optimal number of constructions:

$$\text{diversity} = \frac{\#\{\text{model-generated constructions}\}}{\#\{\text{ShapeWorld-generated constructions}\}}$$

Language constructions here correspond to caption patterns which only record whether an object is described by shape (e.g., *"square"*), colour (e.g., *"red shape"*) or shape and colour (e.g., *"red square"*). So the statement *"A square is red."* and *"A circle is blue."* are considered the same, while *"A shape is red."* is different.

6.3.3 Experimental setup

Datasets. The ShapeWorldICE test-suite (short for ShapeWorld image captioning evaluation) provides 'skill tasks' similar to bAbI (Weston et al., 2015), focusing on (so far) four types of captions: EXISTENTIAL, SPATIAL-EXPLICIT, NUMBERS and QUANTIFIERS (see section 5.3). Each dataset variant consists of a training set of around 200k instances, plus a withheld validation and test set of each 4,096 instances. Training instances consist of an image and a caption, whereas validation/test instances include ten reference captions, which is important for BLEU/SPICE score calculation (here: BLEU-4, up to 4-grams, with uniform weights). Model hyperparameters are tuned using the validation split, and results are reported for the test split only.

Models. We experiment with two image captioning models: the Show&Tell model (Vinyals et al., 2015) and the LRCN_{1u} model (Donahue et al., 2015). Both models follow the basic encoder-decoder architectural pattern of image2seq, using a Inception v3 encoder (Szegedy et al., 2016) to condense the visual information, which in turn conditions an LSTM decoder (Hochreiter and Schmidhuber, 1997) to generate a natural language caption. The main difference between the two models is the way they condition the decoder: Show&Tell feeds the image embedding as 'zeroth word', while LRCN_{1u} concatenates the image features with the LSTM input at every step. The encoder, pretrained on object recognition, and decoder, with word embeddings randomly initialised, are jointly optimised for the standard cross-entropy sequence loss. We train models end-to-end for 100k iterations with a batch size of 64, using Adam optimisation (Kingma and Ba, 2015) with a learning rate of 0.001.

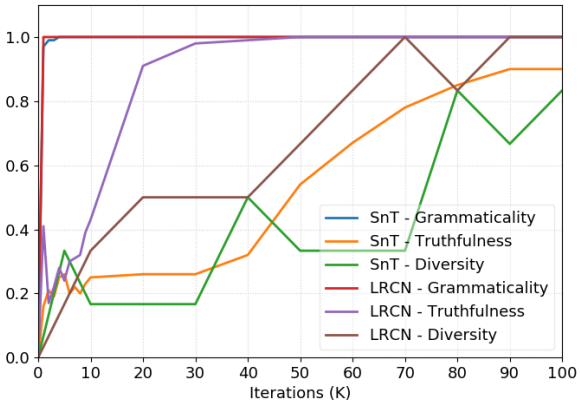


Figure 6.12: GTD performance (*y-axis*) comparison of the Show&Tell model (SnT) and the LRCN_{1u} model (LRCN) on EXISTENTIAL data.

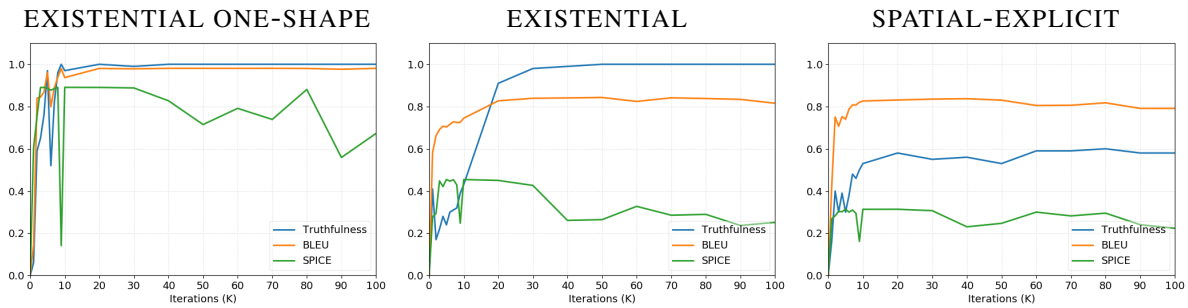


Figure 6.13: Learning curves for LRCN_{1u} on three different datasets (*y-axis*: metric score). BLEU and SPICE denote the average BLEU-4 and SPICE scores across the test split, respectively.

6.3.4 Results

LRCN_{1u} produces more valid and diverse captions. Figure 6.12 illustrates that, while both Show&Tell and the LRCN_{1u} model produce grammatical sentences early on, only the latter learns to consistently generate valid 'true' captions, achieving a truthfulness score of 1.0 halfway through training, whereas Show&Tell only reaches around 0.9. The output of LRCN_{1u} is also consistently more diverse than the captions produced by Show&Tell. We observed similar results on the other datasets, and thus decided to focus on the LRCN_{1u} architecture in the following.

No relation between BLEU, SPICE and truthfulness. Comparing the truthfulness metric with the BLEU and SPICE score on different datasets shows that there is no clear relation between the values of the former and the latter two, as shown in figure 6.13. On the one hand, whereas truthfulness reaches 1.0 early on in training for the first two experiments, both BLEU and SPICE scores differ by roughly 0.2 and $0.4 - 0.5$, respectively. On the other hand, BLEU and SPICE are virtually the same for the second and third experiments, but the truthfulness metric decreases substantially by around 0.4. Importantly, BLEU and SPICE scores change in accordance with each other, despite by different absolute values, suggesting that they capture similar aspects about the output captions. With respect to output captions being valid descriptions of the visual content, these metrics are thus consistently misleading.

Performance on other datasets deteriorates. Figure 6.14 presents the experimental results for LRCN_{1u} on all ShapeWorldICE datasets. The model only reaches a truthfulness score of around $0.45 - 0.6$ when trained on the SPATIAL-EXPLICIT, NUMBERS or QUANTIFIERS (in order of decreasing performance). Since the model does not see any existential statements during training, it only learns to produce the same kinds of more complex captions, which it does not succeed in, despite consistently generating grammatically correct statements. Note, however, that 0.5 is not chance level here, as the space of valid captions for an image is still only a small subset of the space of grammatical statements. In the case of spatial relations, we ran an additional experiment on simplified data where images only consist of two objects. While not perfect, the

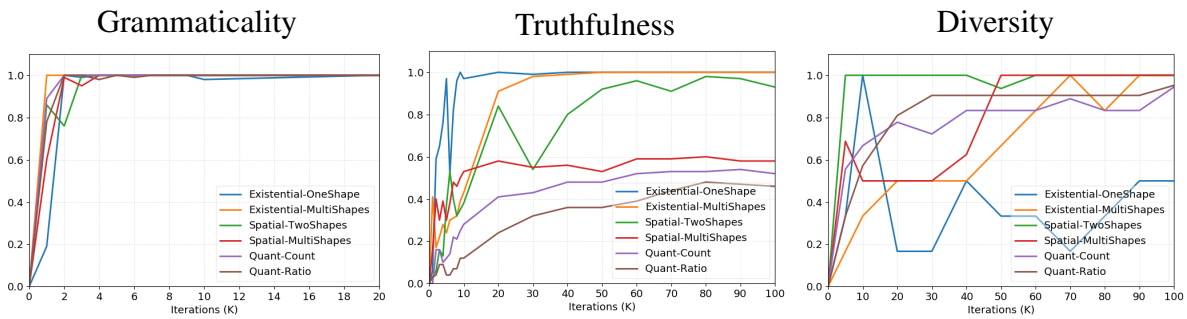


Figure 6.14: GTD performance of LRCN_{1u} on different ShapeWorldICE datasets. Grammaticality is only reported for the first 20k training iterations, as it stays at 100% afterwards.

model achieves a truthfulness performance of more than 0.9, indicating that spatial relations can, in principle, be learned by the model. Another interesting observation is that EXISTENTIAL ONE-SHAPE is the only dataset for which the model fails to produce sufficiently diverse output, despite achieving a perfect truthfulness score.

6.3.5 Conclusion

Evaluation metrics are required as a proxy to measure performance on a task. As such, they should ideally reflect basic requirements for the underlying application. In this work, we propose the GTD evaluation framework as a supplement to standard image captioning evaluation, which explicitly measures grammaticality, truthfulness and diversity. Based on artificial diagnostic captioning datasets, we have evaluated performance of an image captioning model in detail with respect to these metrics. Most importantly, our experiments show that existing metrics like BLEU and SPICE can be completely unrelated to the basic requirements assessed by GTD: they suggest differences where our metrics find none, and report similar behaviour when we confirm substantially lower agreement of captions with visual content. We hope that the GTD framework will enable more insightful image captioning evaluation, and inspire similar approaches to a more accurate assessment of model performance for other generative tasks.

6.4 Other applications of ShapeWorld

Above all, this chapter illustrates how a configurable data simulator like ShapeWorld paves the way to investigate a wide range of different research questions. This includes, for instance, exploring the interplay between data distribution and learning process, borrowing from the rich tradition of behavioural psychology research, or extending the evaluation approach to related tasks like image captioning. The possibilities do not stop here, as other work shows for which ShapeWorld data has also proven to enable interesting experimental evaluation. In the following, two additional avenues of research are briefly introduced.

Natural language provides implicit supervision for structured representations. For many tasks, it is hard to clearly identify the most appropriate representations and abstractions, let alone provide explicit supervision to train a machine learning model. However, natural language acts as an effective and versatile medium for information, which implicitly captures structure useful for a variety of tasks. Inspired by this observation, Andreas et al. (2018) use natural language as a task-independent pretraining step to impose its structure and to bias subsequent learning towards a 'linguistic parametrisation'. One of their illustrative applications is image classification, where they leverage ShapeWorld to produce image data with accompanying in-/valid descriptions. Nash et al. (2018) share the motivation of natural language as implicitly encouraging disentangled representations and, based on that, introduce the Generative Entity Network, a generative model which jointly produces an image and natural language descriptions from a set of latent entities. To confirm superior performance of their model, they use ShapeWorld to produce image data with accompanying partial language descriptions. Generally, a strength of ShapeWorld for these applications is the ability to provide, on the one hand, multiple descriptions for a single image and, on the other hand, invalid statements as contrastive examples.

Signalling games and the emergence of language communication. One potential reason for the emergence of language is the need for communication to achieve a collaborative goal. Signalling games between two agents simulate such a setup: one agent knows the target among a set of objects, the other is required to identify it, and both can communicate discrete symbols to solve this cooperative task. Lars Hulstaert worked on such signalling games using ShapeWorld data as part of his MPhil thesis project in 2016/17, entitled *"Emergence of communication in visually-grounded signalling games"*, which I proposed and co-supervised together with Ann Copestake as main supervisor. Similarly, Graesser et al. (2019) studied language emergence in referential games based on ShapeWorld data. Since the agents are free to develop any form of communication, it is important to be able to control the data distribution, on the one hand, to avoid biases which could otherwise be exploited to artificially increase communication efficiency and, on the other hand, since without a well-defined visual space it is hard to conclusively identify specific phenomena of language emergence. Both is possible with a system like ShapeWorld.

Chapter 7

Conclusion

This thesis is based on two assumptions. First, if deep learning models are able to develop genuine understanding of certain problem patterns which – at least on the surface – resembles human behaviour, then it constitutes an independent research endeavour to assess the capabilities and limitations of different architectures, alongside the traditional focus of machine learning research to evaluate their real-world task performance. Second, if the internal processes and representations learned by deep networks are opaque and will remain resistant to complete theoretical analysis, then their assessment is a matter of experimental investigation following the standards of empirical science, that is, by critically assessing hypotheses via carefully controlled experimental setups. I motivated this thesis with the observation that research in the last years has only insufficiently distinguished between application- and capability-focused evaluation and, as a consequence, has often fallen short of these standards.

At its core, my thesis proposes a novel approach to evaluating black box models such as deep neural networks. This methodology reduces the design of behavioural experiments and implementation of hypotheses to the specification of appropriate data within a configurable data simulation framework, and frames empirical investigation as an incremental process of 'unit-testing': a series of targeted abstract tests with unambiguous outcomes. Such a series of tests constructs an argument for why a model does – or does not – convincingly behave according to the analysed hypothesis. The approach consequently addresses the problem of explainability and interpretability in the context of deep learning, however, not in form of an intrinsic characterisation like a mathematical analysis or formal guarantee, but by means of an extrinsic behavioural assessment, similar to how psychology investigates human decision making – or how software engineering tests correct program functioning.

For its practical contribution, my thesis introduces the ShapeWorld framework, a configurable simulator for visually grounded language data in an abstract domain. I presented a detailed comparative analysis of a set of state-of-the-art visual question answering models, and conducted a series of investigations of other use cases for the ShapeWorld framework. However, my experiments merely scratched the surface of the breadth and depth of potential analyses around

algorithmic capabilities of network modules and behavioural studies of decision making. I want to reiterate that the ShapeWorld system and the various experiments – in addition to being contributions in their own right – are in particular supposed to illustrate the evaluation methodology based on configurable data simulators and unit-testing principles. Consequently, I hope that my work does not just inspire further experiments using ShapeWorld to assess visual question answering models, but also the creation of similar simulator frameworks for other tasks in natural language processing and beyond, as I believe they form an ideal testbed for in-depth capability-focused evaluation of deep learning models.

To conclude, I want to put the content of my thesis into context with respect to three higher-level aspects which I think are most relevant for its contribution to machine learning research going forward: the quest for explainable AI, the emerging science of data generation, and the need for a data toolbox.

The quest for explainable AI. Researchers as well as the public are increasingly concerned about the interpretability of modern machine learning. However, what is required for a satisfactory explanation of model behaviour is an interesting question in itself (Lipton, 2018). I believe that the methodology proposed in this thesis (section 3.2) constitutes one plausible avenue to address concerns around explainability. More specifically, I argue that, whenever we assume a task to require higher-level 'human-like' decision making, the most rigid approach to guarantee that machine learning models behave as expected is to assess them as if their behaviour were human – in fact, for human-centred tasks like natural language understanding, humans are the only sensible reference and there ultimately is no alternative way of 'solving' the task. However, assessing whether behaviour is comparable to humans goes beyond the mere capability to reproduce human annotations for a set of data points. My experiments illustrate how instead behavioural hypotheses around multi-task/curriculum learning (section 6.1) and visual quantifier verification (section 6.2) can be investigated, to shed light on a model's fundamental mechanisms of data processing and how they compare to human inference. Whether an explanation is convincing depends on the situation and audience, as the history of deep learning itself illustrates: neural networks have been researched and applied for decades, but wide-spread concerns about their interpretability emerged only recently. Nonetheless, I think it can be safely said that benchmark datasets have recently struggled not just to produce satisfying explanations, but also to refine them in the face of stricter scrutiny. I believe that this thesis proposes a more flexible and ultimately more convincing evaluation framework to approach the important problem of model interpretability.

The emerging science of data generation. There has been an increasing number of papers using artificial data or entire simulations as part of machine learning research. I envision data generation for machine learning – in particular language data – to be an important emerging

research field in its own right, particularly since there is a variety of questions for which currently every project (re-)invents their own solution: How to integrate a language component with a world simulator? How to guarantee relevant but unbiased output? What are the trade-offs of alternative approaches? How to make the generation process configurable and extensible? How to make the simulation 'scalable', both in terms of quantity and increasing realism? With the considerations around the ShapeWorld framework and its implementation (chapter 4), I attempt to approach this topic in a systematic way. I expect the *"layers of interpretation"* framework of Bender et al. (2015) to be an important building block in this context: (a) to separate purely linguistic concerns from the application; (b) to introduce an application-specific semantics layer on top of a generic language engine, which annotates objects and events of the simulator; and (c) to guarantee efficiency and scalability by embracing language compositionality. The result of a close integration of world and language simulation is best exemplified by the image captioning project (section 6.3), for which the relevant parsing functionality was a mere 'by-product' of ShapeWorld's ability to generate visually grounded language. Recent 'sim2real' approaches (Tobin et al., 2017) try to bridge the gap to reality with increasingly realistic and flexible simulations, however, integration with language is still in its infancy and offers a range of possibilities for future work on data generation.

The need for a data toolbox. Empirical research is about asking questions which can be verified or falsified experimentally. In the case of deep learning, this corresponds to investigating the effect of either architecture modifications on the model side, or data modifications on the application side. Software libraries like TensorFlow or PyTorch have greatly enhanced our ability as researchers to quickly prototype and share modelling ideas. What I believe is missing – and what my thesis attempts to provide – is a similar toolbox for data creation to support the rapid, cheap and reproducible implementation of application-related hypotheses (section 2.5 and 3.2). Crowdsourcing is doubtlessly a first step in this direction – as the recent explosion of crowdsourced datasets testifies – but it is still comparatively inflexible and expensive, and offers little control over the precise content and quality of data. Besides the advantage of inspiring as well as standardising data prototyping, I expect that a data toolbox would have a 'regularising' effect on model development and experimentation in the community as a whole, ultimately leading to more robust results. On the one hand, researchers would be expected to investigate not just model but also data hyperparameters and, on the other hand, it would be much easier for others to spot obvious flaws and instabilities due to insufficient testing.

To sum up, in this thesis I have argued for an evaluation approach orthogonal to benchmark datasets, which explicitly encourages to ask questions and simultaneously provides the means to implement them. By enabling machine learning researchers to formulate many questions, discard unproductive avenues and iterate quickly, I hope to have contributed to more meaningful progress in the coming years.

Bibliography

Adi, Yossi, Einat Kermany, Yonatan Belinkov, Ofer Lavi and Yoav Goldberg (Apr. 2017). 'Fine-grained Analysis of Sentence Embeddings Using Auxiliary Prediction Tasks'. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France. [Link] (see page 38).

Agrawal, Aishwarya, Dhruv Batra and Devi Parikh (Nov. 2016). 'Analyzing the Behavior of Visual Question Answering Models'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, USA, pp. 1955–1960. [Link] (see page 28, 60).

Agrawal, Aishwarya, Dhruv Batra, Devi Parikh and Aniruddha Kembhavi (June 2018). 'Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA, pp. 4971–4980. [Link] (see page 34, 61).

Agrawal, Aishwarya, Aniruddha Kembhavi, Dhruv Batra and Devi Parikh (2017). *C-VQA: A Compositional Split of the Visual Question Answering (VQA) v1.0 Dataset*. arXiv: 1704.08243 (see page 34, 61).

Anand, Ankesh, Eugene Belilovsky, Kyle Kastner, Hugo Larochelle and Aaron Courville (2018). *Blindfold Baselines for Embodied QA*. arXiv: 1811.05013 (see page 28).

Anderson, Peter, Basura Fernando, Mark Johnson and Stephen Gould (Oct. 2016). 'SPICE: Semantic Propositional Image Caption Evaluation'. In: *Proceedings of the 14th European Conference on Computer Vision (ECCV)*. Amsterdam, The Netherlands, pp. 382–398. [Link] (see page 32, 138).

Andreas, Jacob, Dan Klein and Sergey Levine (June 2018). 'Learning with Latent Language'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 2166–2179. [Link] (see page 144).

Andreas, Jacob, Marcus Rohrbach, Trevor Darrell and Dan Klein (June 2016a). 'Learning to Compose Neural Networks for Question Answering'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. San Diego, CA, USA, pp. 1545–1554. [Link] (see page 59).

— (June 2016b). ‘Neural Module Networks’. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 39–48. [Link] (see page 59, 61).

Antol, Stanislaw, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick and Devi Parikh (Dec. 2015). ‘VQA: Visual Question Answering’. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, pp. 2425–2433. [Link] (see page 28, 30, 34, 37, 57–59, 61, 121, 136).

Arora, Sanjeev, Yingyu Liang and Tengyu Ma (Apr. 2017). ‘A Simple but Tough-to-Beat Baseline for Sentence Embeddings’. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France. [Link] (see page 25).

Arthur, Philip, Graham Neubig and Satoshi Nakamura (Nov. 2016). ‘Incorporating Discrete Translation Lexicons into Neural Machine Translation’. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, USA, pp. 1557–1567. [Link] (see page 21).

Atzmon, Yuval, Jonathan Berant, Vahid Kezami, Amir Globerson and Gal Chechik (2016). *Learning to generalize to new compositions in image understanding*. arXiv: 1608.07639 (see page 34).

Avcu, Enes, Chihiro Shibata and Jeffrey Heinz (2017). *Subregular Complexity and Deep Learning*. arXiv: 1705.05940 (see page 35, 36).

Bai, Shaojie, J. Zico Kolter and Vladlen Koltun (2018). *An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling*. arXiv: 1803.01271 (see page 24).

Baldwin, Timothy, Valia Kordoni and Aline Villavicencio (2009). ‘Prepositions in Applications: A Survey and Introduction to the Special Issue’. In: *Computational Linguistics* 35.2, pp. 119–149. [Link] (see page 27).

Banerjee, Satanjeev and Alon Lavie (June 2005). ‘METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments’. In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, MI, USA, pp. 65–72. [Link] (see page 138).

Barratt, Shane and Rishi Kant Sharma (2018). *A Note on the Inception Score*. arXiv: 1801.01973 (see page 32).

Barrett, David, Felix Hill, Adam Santoro, Ari Morcos and Timothy Lillicrap (July 2018). ‘Measuring abstract reasoning in neural networks’. In: *Proceedings of the 35th International Conference on Machine Learning (ICML)*. Stockholm, Sweden, pp. 511–520. [Link] (see page 36).

Bastings, Joost, Marco Baroni, Jason Weston, Kyunghyun Cho and Douwe Kiela (Nov. 2018). ‘Jump to better conclusions: SCAN both left and right’. In: *Proceedings of the EMNLP*

Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Brussels, Belgium, pp. 47–55. [Link] (see page 36).

Belinkov, Yonatan and Yonatan Bisk (2017). *Synthetic and Natural Noise Both Break Neural Machine Translation*. arXiv: 1711.02173 (see page 21).

Ben-Younes, Hedi, Rémi Cadène, Nicolas Thome and Matthieu Cord (Oct. 2017). 'MUTAN: Multimodal Tucker Fusion for Visual Question Answering'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, pp. 2631–2639. [Link] (see page 59).

Bender, Emily M., Dan Flickinger and Stephan Oepen (Aug. 2002). 'The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-linguistically Consistent Broad-Coverage Precision Grammars'. In: *Proceedings of the COLING Workshop on Grammar Engineering and Evaluation*. Taipei, Taiwan. [Link] (see page 87, 88).

Bender, Emily M., Dan Flickinger, Stephan Oepen, Woodley Packard and Ann Copestake (Apr. 2015). 'Layers of Interpretation: On Grammar and Compositionality'. In: *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*. London, United Kingdom, pp. 239–249. [Link] (see page 80, 147).

Bengio, Yoshua, Jérôme Louradour, Ronan Collobert and Jason Weston (June 2009). 'Curriculum Learning'. In: *Proceedings of the 26th International Conference on Machine Learning (ICML)*. Montreal, Canada, pp. 41–48. [Link] (see page 125).

Bennett, Craig M., Abigail A. Baird, Michael B. Miller and George L. Wolford (2009). 'Neural correlates of interspecies perspective taking in the post-mortem Atlantic Salmon: An argument for multiple comparisons correction'. In: *Journal of Serendipitous and Unexpected Results* 1.1, pp. 1–5. [Link] (see page 33).

Bernardy, Jean-Philippe and Stergios Chatzikyriakidis (2018). *A corpus of precise natural textual entailment problems*. arXiv: 1812.05813 (see page 35).

Bishop, Christopher M. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer Science+Business Media. [Link] (see page 19).

Bisk, Yonatan, Deniz Yuret and Daniel Marcu (June 2016). 'Natural Language Communication with Robots'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. San Diego, CA, USA, pp. 751–761. [Link] (see page 35, 36).

Bordes, Antoine, Nicolas Usunier, Sumit Chopra and Jason Weston (2015). *Large-scale Simple Question Answering with Memory Networks*. arXiv: 1506.02075 (see page 28).

Bowman, Samuel R. (2013). *Can recursive neural tensor networks learn logical reasoning?* arXiv: 1312.6192 (see page 35).

Bowman, Samuel R., Gabor Angeli, Christopher Potts and Christopher D. Manning (Sept. 2015). 'A large annotated corpus for learning natural language inference'. In: *Proceedings of*

the Conference on Empirical Methods in Natural Language Processing (EMNLP). Lisbon, Portugal, pp. 632–642. [Link] (see page 23, 28, 29, 34, 38, 121).

Callison-Burch, Chris, Miles Osborne and Philipp Koehn (Apr. 2006). 'Re-evaluation the Role of BLEU in Machine Translation Research'. In: *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Trento, Italy. [Link] (see page 32).

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm and Noemie Elhadad (Aug. 2015). 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission'. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Sydney, Australia, pp. 1721–1730. [Link] (see page 21).

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan L. Boyd-graber and David M. Blei (2009). 'Reading Tea Leaves: How Humans Interpret Topic Models'. In: *Proceedings of the 23rd International Conference on Neural Information Processing Systems (NIPS)*. Vancouver, Canada, pp. 288–296. [Link] (see page 32).

Chao, Wei-Lun, Hexiang Hu and Fei Sha (June 2018). 'Being Negative but Constructively: Lessons Learnt from Creating Better Visual Question Answering Datasets'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 431–441. [Link] (see page 37, 61).

Chattopadhyay, Prithvijit, Ramakrishna Vedantam, Ramprasaath R. Selvaraju, Dhruv Batra and Devi Parikh (June 2017). 'Counting Everyday Objects in Everyday Scenes'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, pp. 4428–4437. [Link] (see page 137).

Chatzikyriakidis, Stergios, Robin Cooper, Simon Dobnik and Staffan Larsson (Aug. 2017). 'An overview of Natural Language Inference Data Collection: The way forward?' In: *Proceedings of the ACL Workshop on Computing Natural Language Inference*. Vancouver, Canada. [Link] (see page 35).

Chen, Danqi, Jason Bolton and Christopher D. Manning (Aug. 2016). 'A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task'. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Berlin, Germany, pp. 2358–2367. [Link] (see page 24).

Chiticariu, Laura, Yunyao Li and Frederick R. Reiss (Oct. 2013). 'Rule-Based Information Extraction is Dead! Long Live Rule-Based Information Extraction Systems!' In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Seattle, WA, USA, pp. 827–832. [Link] (see page 41).

Cho, Kyunghyun, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio (Oct. 2014). 'Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation'. In: *Proceedings of the Con-*

ference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar, pp. 1724–1734. [Link] (see page 94).

Chrabaszcz, Patryk, Ilya Loshchilov and Frank Hutter (July 2018). 'Back to Basics: Benchmarking Canonical Evolution Strategies for Playing Atari'. In: *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*. Stockholm, Sweden, pp. 1419–1426. [Link] (see page 25).

Cirik, Volkan, Louis-Philippe Morency and Taylor Berg-Kirkpatrick (June 2018). 'Visual Referring Expression Recognition: What Do Systems Actually Learn?' In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 781–787. [Link] (see page 28, 31, 38).

Clark, Peter, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick and Oyvind Tafjord (2018). *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. arXiv: 1803.05457 (see page 35).

Conneau, Alexis, Germán Kruszewski, Guillaume Lample, Loïc Barrault and Marco Baroni (July 2018). 'What you can cram into a single \$&!#* vector: Probing sentence embeddings for linguistic properties'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 2126–2136. [Link] (see page 25, 38).

Cooper, Robin, Dick Crouch, Jan van Eijck, Chris Fox, Johan van Genabith, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio and Steve Pulman (1996). *Using the framework*. Technical Report LRE 62-051 D-16. The FraCaS Consortium. [Link] (see page 35).

Copestake, Ann (Apr. 2009). 'Slacker Semantics: Why Superficiality, Dependency and Avoidance of Commitment can be the Right Way to Go'. In: *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Athens, Greece, pp. 1–9. [Link] (see page 78, 79).

Copestake, Ann, Guy Emerson, Michael W. Goodman, Matic Horvat, Alexander Kuhnle and Ewa Muszyńska (May 2016). 'Resources for Building Applications with Dependency Minimal Recursion Semantics'. In: *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC)*. Portorož, Slovenia, pp. 1240–1247. [Link] (see page 17, 78, 82).

Copestake, Ann, Dan Flickinger, Carl Pollard and Ivan A. Sag (2005). 'Minimal Recursion Semantics: An Introduction'. In: *Research on Language and Computation* 3.2, pp. 281–332. [Link] (see page 78).

Daniluk, Michal, Tim Rocktäschel, Johannes Welbl and Sebastian Riedel (2017). *Frustratingly Short Attention Spans in Neural Language Modeling*. arXiv: 1702.04521 (see page 24).

Dasgupta, Ishita, Demi Guo, Andreas Stuhlmüller, Samuel J. Gershman and Noah D. Goodman (2018). *Evaluating Compositionality in Sentence Embeddings*. arXiv: 1802.04302 (see page 31, 38).

Demšar, Janez (July 2008). 'On the Appropriateness of Statistical Tests in Machine Learning'. In: *Proceedings of the ICML Workshop on Evaluation Methods for Machine Learning*. Helsinki, Finland, pp. 1–4. [Link] (see page 33).

Devlin, Jacob, Ming-Wei Chang, Kenton Lee and Kristina Toutanova (June 2019). 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Minneapolis, MN, USA, pp. 4171–4186. [Link] (see page 39).

Devlin, Jacob, Saurabh Gupta, Ross Girshick, Margaret Mitchell and C. Lawrence Zitnick (2015). *Exploring nearest neighbor approaches for image captioning*. arXiv: 1505.04467 (see page 24).

Ding, Nan, Sebastian Goodman, Fei Sha and Radu Soricut (2016). *Understanding Image and Text Simultaneously: a Dual Vision-Language Machine Comprehension Task*. arXiv: 1612.07833 (see page 37).

Donahue, Jeff, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Trevor Darrell and Kate Saenko (June 2015). 'Long-term recurrent convolutional networks for visual recognition and description'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA, pp. 2625–2634. [Link] (see page 141).

Dubossarsky, Haim, Eitan Grossman and Daphna Weinshall (Nov. 2018). 'Coming to Your Senses: on Controls and Evaluation Sets in Polysemy Research'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 1732–1740. [Link] (see page 24).

Dukes, Kais (Aug. 2014). 'SemEval-2014 Task 6: Supervised Semantic Parsing of Robotic Spatial Commands'. In: *Proceedings of the COLING Workshop on Semantic Evaluation*. Dublin, Ireland, pp. 45–53. [Link] (see page 35).

Dumoulin, Vincent, Ethan Perez, Nathan Schucher, Florian Strub, Harm de Vries, Aaron Courville and Yoshua Bengio (2018). 'Feature-wise transformations'. In: *Distill*. [Link] (see page 62, 63).

Elliott, Desmond and Frank Keller (June 2014). 'Comparing Automatic Evaluation Measures for Image Description'. In: *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*. Baltimore, MD, USA, pp. 452–457. [Link] (see page 32).

Elman, Jeffrey L. (1993). 'Learning and development in neural networks: the importance of starting small'. In: *Cognition* 48.1, pp. 71–99. [Link] (see page 125).

Ettinger, Allyson, Sudha Rao, Hal Daumé III and Emily M. Bender (Sept. 2017). 'Towards Linguistically Generalizable NLP Systems: A Workshop and Shared Task'. In: *Proceedings of*

the EMNLP Workshop on Building Linguistically Generalizable NLP Systems. Copenhagen, Denmark, pp. 1–10. [Link] (see page 35).

Evans, Richard, David Saxton, David Amos, Pushmeet Kohli and Edward Grefenstette (2018). *Can Neural Networks Understand Logical Entailment?* arXiv: 1802.08535 (see page 35, 36).

Eysenbach, Benjamin, Carl Vondrick and Antonio Torralba (2016). *Who is Mistaken?* arXiv: 1612.01175 (see page 39).

Fan, Zhenzhen, Sanghoun Song and Francis Bond (July 2015). 'An HPSG-based Shared-Grammar for the Chinese Languages: ZHONG [—]'. In: *Proceedings of the ACL Workshop on Grammar Engineering Across Frameworks*. Beijing, China, pp. 17–24. [Link] (see page 87, 88).

Fang, Yimai, Haoyue Zhu, Ewa Muszyńska, Alexander Kuhnle and Simone Teufel (Dec. 2016). 'A Proposition-Based Abstractive Summariser'. In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*. Osaka, Japan, pp. 567–578. [Link] (see page 17).

Feng, Shi, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez and Jordan Boyd-Graber (Nov. 2018). 'Pathologies of Neural Models Make Interpretations Difficult'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 3719–3728. [Link] (see page 22).

Flickinger, Dan (2000). 'On Building a More Efficient Grammar by Exploiting Types'. In: *Natural Language Engineering* 6.1, pp. 15–28. [Link] (see page 78).

— (2011). 'Accuracy vs. Robustness in Grammar Engineering'. In: *Language from a Cognitive Perspective: Grammar, Usage, and Processing*. Ed. by Emily M. Bender and Jennifer E. Arnold. CSLI Publications, pp. 31–50. [Link] (see page 78).

Flickinger, Dan, Emily M. Bender and Stephan Oepen (May 2014). 'Towards an Encyclopedia of Compositional Semantics: Documenting the Interface of the English Resource Grammar'. In: *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*. Reykjavik, Iceland, pp. 875–881. [Link] (see page 78).

Fokkens, Antske, Marieke van Erp, Marten Postma, Ted Pedersen, Piek Vossen and Nuno Freire (Aug. 2013). 'Offspring from Reproduction Problems: What Replication Failure Teaches Us'. In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Sofia, Bulgaria, pp. 1691–1701. [Link] (see page 23).

Fukui, Akira, Dong Huk Park, Daylen Yang, Anna Rohrbach, Trevor Darrell and Marcus Rohrbach (Nov. 2016). 'Multimodal Compact Bilinear Pooling for Visual Question Answering and Visual Grounding'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, USA, pp. 457–468. [Link] (see page 59).

Gao, Haoyuan, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang and Wei Xu (Dec. 2015). 'Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question Answering'. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, pp. 2296–2304. [Link] (see page 57–59, 88).

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jenn Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (2018). *Datasheets for Datasets*. arXiv: 1803.09010 (see page 55).

Geiger, Atticus, Ignacio Cases, Lauri Karttunen and Christopher Potts (2018). *Stress-Testing Neural Models of Natural Language Inference with Multiply-Quantified Sentences*. arXiv: 1810.13033 (see page 38, 39).

Geirhos, Robert, P. Rubisch, C. Michaelis, Matthias Bethge, Felix A. Wichmann and W. Brendel (May 2019). 'ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness'. In: *Proceedings of the 7th International Conference on Learning Representations (ICLR)*. New Orleans, LA, USA. [Link] (see page 39, 118).

Geirhos, Robert, Carlos R. M. Temme, Jonas Rauber, Heiko H. Schütt, Matthias Bethge and Felix A. Wichmann (Dec. 2018). 'Generalisation in humans and deep neural networks'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, pp. 7538–7550. [Link] (see page 38).

Geman, Donald, Stuart Geman, Neil Hallonquist and Laurent Younes (2015). 'Visual Turing test for computer vision systems'. In: *Proceedings of the National Academy of Sciences* 112.12, pp. 3618–3623. [Link] (see page 57, 58).

Gers, F. A. and J. Schmidhuber (2001). 'LSTM Recurrent Networks Learn Simple Context-free and Context-sensitive Languages'. In: *IEEE Transactions on Neural Networks* 12.6, pp. 1333–1340. [Link] (see page 35).

Glockner, Max, Vered Shwartz and Yoav Goldberg (July 2018). 'Breaking NLI Systems with Sentences that Require Simple Lexical Inferences'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 650–655. [Link] (see page 37).

Glorot, Xavier, Antoine Bordes and Yoshua Bengio (Apr. 2011). 'Deep Sparse Rectifier Neural Networks'. In: *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL, USA, pp. 315–323. [Link] (see page 93).

Goldberg, Yoav (2019). *Assessing BERT's Syntactic Abilities*. arXiv: 1901.05287 (see page 39).

Goyal, Yash, Tejas Khot, Douglas Summers-Stay, Dhruv Batra and Devi Parikh (June 2017). 'Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, pp. 6325–6334. [Link] (see page 28, 34, 60, 61).

Graesser, Laura, Kyunghyun Cho and Douwe Kiela (2019). *Emergent Linguistic Phenomena in Multi-Agent Communication Games*. arXiv: 1901.08706 (see page 144).

Gururangan, Suchin, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman and Noah A. Smith (June 2018). 'Annotation Artifacts in Natural Language Inference Data'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 107–112. [Link] (see page 28, 29, 51).

Hackl, Martin (2009). 'On the grammar and processing of proportional quantifiers: most versus more than half'. In: *Natural Language Semantics* 17.1, pp. 63–98. [Link] (see page 129).

Halevy, Alon, Peter Norvig and Fernando Pereira (2009). 'The Unreasonable Effectiveness of Data'. In: *IEEE Intelligent Systems* 24.2, pp. 8–12. [Link] (see page 20, 121).

He, Kaiming, Xiangyu Zhang, Shaoqing Ren and Jian Sun (June 2016). 'Deep Residual Learning for Image Recognition'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 770–778. [Link] (see page 42, 93).

Henderson, Peter, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup and David Meger (Feb. 2018). 'Deep Reinforcement Learning That Matters'. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA, pp. 3207–3214. [Link] (see page 23, 40).

Hendrycks, Dan and Thomas G. Dietterich (2018). *Benchmarking Neural Network Robustness to Common Corruptions and Surface Variations*. arXiv: 1807.01697 (see page 38).

Hermann, Karl Moritz, Tomáš Kočiský, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman and Phil Blunsom (Dec. 2015). 'Teaching Machines to Read and Comprehend'. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, pp. 1693–1701. [Link] (see page 29).

Hill, Felix, Antoine Bordes, Sumit Chopra and Jason Weston (May 2016). 'The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations'. In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico. [Link] (see page 29).

Hochreiter, Sepp and Jürgen Schmidhuber (1997). 'Long Short-Term Memory'. In: *Neural Computation* 9.8, pp. 1735–1780. [Link] (see page 94, 141).

Hodosh, Micah and Julia Hockenmaier (Aug. 2016). 'Focused Evaluation for Image Description with Binary Forced-Choice Tasks'. In: *Proceedings of the ACL Workshop on Vision and Language*. Berlin, Germany, pp. 19–28. [Link] (see page 37).

Hodosh, Micah, Peter Young and Julia Hockenmaier (2013). 'Framing Image Description As a Ranking Task: Data, Models and Evaluation Metrics'. In: *Journal of Artificial Intelligence Research* 47.1, pp. 853–899. [Link] (see page 37).

Horn, Grant van and Pietro Perona (2017). *The Devil is in the Tails: Fine-grained Classification in the Wild*. arXiv: 1709.01450 (see page 29, 30).

Hoshen, Dokhyam and Michael Werman (2017). *IQ of Neural Networks*. arXiv: 1710.01692 (see page 36).

Hu, Ronghang, Jacob Andreas, Marcus Rohrbach, Trevor Darrell and Kate Saenko (Oct. 2017). 'Learning to Reason: End-to-End Module Networks for Visual Question Answering'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, pp. 804–813. [Link] (see page 62, 88).

Hudson, Drew A. and Christopher D. Manning (May 2018). 'Compositional Attention Networks for Machine Reasoning'. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada. [Link] (see page 62, 137).

Hutson, Matthew (2018). 'Artificial intelligence faces reproducibility crisis'. In: *Science* 359.6377, pp. 725–726. [Link] (see page 20, 40).

Ilievski, Ilija, Shuicheng Yan and Jiashi Feng (2016). *A Focused Dynamic Attention Model for Visual Question Answering*. arXiv: 1604.01485 (see page 59).

Ioannidis, John P. A. (2005). 'Why Most Published Research Findings Are False'. In: *PLOS Medicine* 2.8. [Link] (see page 33, 39–41).

Ioffe, Sergey and Christian Szegedy (July 2015). 'Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift'. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille, France, pp. 448–456. [Link] (see page 42, 93).

Isabelle, Pierre, Colin Cherry and George Foster (Sept. 2017). 'A Challenge Set Approach to Evaluating Machine Translation'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark, pp. 2486–2496. [Link] (see page 35).

Jabri, Allan, Armand Joulin and Laurens van der Maaten (Oct. 2016). 'Revisiting Visual Question Answering Baselines'. In: *Proceedings of the 14th European Conference on Computer Vision (ECCV)*. Amsterdam, The Netherlands, pp. 727–739. [Link] (see page 25, 60, 63).

Jia, Robin and Percy Liang (Sept. 2017). 'Adversarial Examples for Evaluating Reading Comprehension Systems'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark, pp. 2021–2031. [Link] (see page 22, 30, 38).

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick and Ross Girshick (June 2017a). 'CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI, USA, pp. 1988–1997. [Link] (see page 36, 61, 70, 80, 91, 93, 96, 98–100, 136).

Johnson, Justin, Bharath Hariharan, Laurens van der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick and Ross Girshick (Oct. 2017b). 'Inferring and Executing Programs for Visual

Reasoning'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, pp. 3008–3017. [Link] (see page 62, 88, 93, 95, 96, 100).

Jozefowicz, Rafal, Wojciech Zaremba and Ilya Sutskever (July 2015). 'An Empirical Exploration of Recurrent Network Architectures'. In: *Proceedings of the 32nd International Conference on Machine Learning (ICML)*. Lille, France, pp. 2342–2350. [Link] (see page 24).

Jumelet, Jaap and Dieuwke Hupkes (Nov. 2018). 'Do Language Models Understand Anything? On the Ability of LSTMs to Understand Negative Polarity Items'. In: *Proceedings of the EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium, pp. 222–231. [Link] (see page 39).

Kafle, Kushal and Christopher Kanan (Oct. 2017a). 'An Analysis of Visual Question Answering Algorithms'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, pp. 1983–1991. [Link] (see page 34, 60, 61).

— (2017b). 'Visual question answering: Datasets, algorithms, and future challenges'. In: *Computer Vision and Image Understanding* 163, pp. 3–20. [Link] (see page 30, 60).

Kahou, Samira Ebrahimi, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler and Yoshua Bengio (May 2018). 'FigureQA: An Annotated Figure Dataset for Visual Reasoning'. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada. [Link] (see page 67).

Kaushik, Divyansh and Zachary C. Lipton (Nov. 2018). 'How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 5010–5015. [Link] (see page 29).

Kazemi, Vahid and Ali Elqursh (2017). *Show, Ask, Attend, and Answer: A Strong Baseline For Visual Question Answering*. arXiv: 1704.03162 (see page 25, 59, 60, 63).

Kilickaya, Mert, Aykut Erdem, Nazli Ikizler-Cinbis and Erkut Erdem (Apr. 2017). 'Re-evaluating Automatic Metrics for Image Captioning'. In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Valencia, Spain, pp. 199–209. [Link] (see page 32).

Kim, Jin-Hwa, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha and Byoung-Tak Zhang (Apr. 2017). 'Hadamard Product for Low-rank Bilinear Pooling'. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France. [Link] (see page 59).

Kim, Junkyung, Matthew Ricci and Thomas Serre (2018). 'Not-So-CLEVR: learning same-different relations strains feedforward neural networks'. In: *Interface Focus* 8.4. [Link] (see page 36).

Kingma, Diederik P. and Jimmy Ba (May 2015). 'Adam: A Method for Stochastic Optimization'. In: *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*. San Diego, CA, USA. [Link] (see page 95, 141).

Király, Franz J., Bilal Mateen and Raphael Sonabend (2018). *NIPS – Not Even Wrong? A Systematic Review of Empirically Complete Demonstrations of Algorithmic Effectiveness in the Machine Learning and Artificial Intelligence Literature*. arXiv: 1812.07519 (see page 33, 40).

Kornblith, Simon, Jon Shlens and Quoc V. Le (June 2019). 'Do better ImageNet models transfer better?' In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA. [Link] (see page 23).

Krizhevsky, Alex, Ilya Sutskever and Geoffrey E. Hinton (Dec. 2012). 'ImageNet Classification with Deep Convolutional Neural Networks'. In: *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, pp. 1097–1105. [Link] (see page 19).

Kuhnle, Alexander (2016). *GraphLang: DMRS graph description language*. [Link] (see page 81, 82).

Kuhnle, Alexander and Ann Copestake (2017). *ShapeWorld – A new test methodology for multimodal language understanding*. arXiv: 1704.04517 (see page 17, 61).

— (June 2018). 'Deep learning evaluation using deep linguistic processing'. In: *Proceedings of the NAACL Workshop on Generalization in the Age of Deep Learning*. New Orleans, LA, USA, pp. 17–23. [Link] (see page 17).

— (Aug. 2019a). 'The meaning of "most" for visual question answering models'. In: *Proceedings of the ACL Workshop on BlackboxNLP*. Florence, Italy. [Link] (see page 17, 126, 137).

— (Dec. 2019b). 'What is needed for simple spatial language capabilities in VQA?' In: *Proceedings of the NeurIPS Workshop on Visually Grounded Interaction and Language*. Vancouver, Canada. [Link] (see page 17, 114).

Kuhnle, Alexander, Huiyuan Xie and Ann Copestake (Sept. 2018). 'How Clever Is the FiLM Model, and How Clever Can it Be?' In: *Proceedings of the ECCV Workshops*. Munich, Germany, pp. 162–172. [Link] (see page 17, 120).

Kumar, Ankit, Ozan Irsoy, Peter Ondruska, Mohit Iyyer, James Bradbury, Ishaan Gulrajani, Victor Zhong, Romain Paulus and Richard Socher (June 2016). 'Ask Me Anything: Dynamic Memory Networks for Natural Language Processing'. In: *Proceedings of the 33rd International Conference on Machine Learning (ICML)*. New York, NY, USA, pp. 1378–1387. [Link] (see page 59).

Labutov, Igor, Bishan Yang, Anusha Prakash and Amos Azaria (July 2018). 'Multi-Relational Question Answering from Narratives: Machine Reading and Reasoning in Simulated Worlds'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 833–844. [Link] (see page 36).

Lake, Brenden M. and Marco Baroni (July 2018). 'Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks'. In: *Proceedings*

of the 35th International Conference on Machine Learning (ICML). Stockholm, Sweden, pp. 2879–2888. [Link] (see page 36).

Langley, Pat (2011). 'The Changing Science of Machine Learning'. In: *Machine Learning* 82.3, pp. 275–279. [Link] (see page 22, 41).

LeCun, Yann, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard and L. D. Jackel (1989). 'Backpropagation Applied to Handwritten Zip Code Recognition'. In: *Neural Computation* 1.4, pp. 541–551. [Link] (see page 93).

Lee, Moontae, Xiaodong He, Wen-Tau Yih, Jianfeng Gao, Li Deng and Paul Smolensky (May 2016). 'Reasoning in Vector Space: An Exploratory Study of Question Answering'. In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico. [Link] (see page 30).

Levesque, Hector J. (2014). 'On Our Best Behaviour'. In: *Artificial Intelligence* 212.1, pp. 27–35. [Link] (see page 37, 41).

Levesque, Hector J., Ernest Davis and Leora Morgenstern (June 2012). 'The Winograd Schema Challenge'. In: *Proceedings of the 13th International Conference on Principles of Knowledge Representation and Reasoning*. Rome, Italy, pp. 552–561. [Link] (see page 30, 35, 37).

Levy, Omer, Steffen Remus, Chris Biemann and Ido Dagan (June 2015). 'Do Supervised Distributional Methods Really Learn Lexical Inference Relations?' In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Denver, CO, USA, pp. 970–976. [Link] (see page 24, 29).

Li, Yuanpeng, Yi Yang, Jianyu Wang and Wei Xu (2018). *Zero-Shot Transfer VQA Dataset*. arXiv: `1811.00692` (see page 34, 61).

Lidz, Jeffrey, Paul Pietroski, Justin Halberda and Tim Hunter (2011). 'Interface transparency and the psychosemantics of most'. In: *Natural Language Semantics* 19.3, pp. 227–256. [Link] (see page 130).

Lin, Chin-Yew (July 2004). 'ROUGE: A Package for Automatic Evaluation of Summaries'. In: *Proceedings of the ACL Workshop on Text Summarization Branches Out*. Barcelona, Spain, pp. 74–81. [Link] (see page 138).

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár and C. Lawrence Zitnick (Sept. 2014). 'Microsoft COCO: Common Objects in Context'. In: *Proceedings of the 13th European Conference on Computer Vision (ECCV)*. Zurich, Switzerland, pp. 740–755. [Link] (see page 29, 30, 34, 37, 51, 58).

Linzen, Tal, Emmanuel Dupoux and Yoav Goldberg (2016). 'Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies'. In: *Transactions of the Association for Computational Linguistics* 4, pp. 521–535. [Link] (see page 30, 34–36).

Lipton, Zachary C. (2018). 'The Mythos of Model Interpretability'. In: *Queue* 16.3, pp. 31–57. [Link] (see page 146).

Lipton, Zachary C. and Jacob Steinhardt (2018). *Troubling Trends in Machine Learning Scholarship*. arXiv: 1807.03341 (see page 20, 40, 41).

Liska, Adam, Germán Kruszewski and Marco Baroni (2018). *Memorize or generalize? Searching for a compositional RNN in a haystack*. arXiv: 1802.06467 (see page 36).

Litkowski, Ken and Orin Hargraves (Apr. 2006). 'Coverage and Inheritance in the Preposition Project'. In: *Proceedings of the ACL Workshop on Prepositions*. Trento, Italy, pp. 37–44. [Link] (see page 27).

Liu, Chia-Wei, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin and Joelle Pineau (Nov. 2016). 'How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, USA, pp. 2122–2132. [Link] (see page 32).

Lu, Jiasen, Xiao Lin, Dhruv Batra and Devi Parikh (2015). *Deeper LSTM and normalized CNN Visual Question Answering model*. [Link] (see page 25, 60, 63, 92).

Lu, Jiasen, Jianwei Yang, Dhruv Batra and Devi Parikh (Dec. 2016). 'Hierarchical Question-image Co-attention for Visual Question Answering'. In: *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS)*. Barcelona, Spain, pp. 289–297. [Link] (see page 59).

Lučić, Mario, Karol Kurach, Marcin Michalski, Sylvain Gelly and Olivier Bousquet (Dec. 2018). 'Are GANs Created Equal? A Large-Scale Study'. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada. [Link] (see page 24, 32, 36).

Ma, Lin, Zhengdong Lu and Hang Li (Feb. 2016). 'Learning to Answer Questions from Image Using Convolutional Neural Network'. In: *Proceedings of the 30th AAAI Conference on Artificial Intelligence*. Phoenix, AZ, USA, pp. 3567–3573. [Link] (see page 59).

Madhyastha, Pranava, Josiah Wang and Lucia Specia (July 2019). 'VIFIDEL: Evaluating the Visual Fidelity of Image Descriptions'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy, pp. 6539–6550. [Link] (see page 139).

Mahendru, Aroma, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra and Stefan Lee (Sept. 2017). 'The Promise of Premise: Harnessing Question Premises in Visual Question Answering'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Copenhagen, Denmark, pp. 926–935. [Link] (see page 37, 61).

Malinowski, Mateusz and Carl Doersch (Sept. 2018). 'The Visual QA Devil in the Details: The Impact of Early Fusion and Batch Norm on CLEVR'. In: *Proceedings of the ECCV Workshops*. Munich, Germany. [Link] (see page 63, 92, 93, 95, 98, 99).

Malinowski, Mateusz and Mario Fritz (Dec. 2014a). 'A Multi-world Approach to Question Answering About Real-world Scenes Based on Uncertain Input'. In: *Proceedings of the*

28th International Conference on Neural Information Processing Systems (NIPS). Montreal, Canada, pp. 1682–1690. [Link] (see page 57, 58, 61).

— (Dec. 2014b). 'Towards a Visual Turing Challenge'. In: *Proceedings of the NIPS Workshop on Learning Semantics*. Montreal, Canada. [Link] (see page 57).

Malinowski, Mateusz, Marcus Rohrbach and Mario Fritz (Dec. 2015). 'Ask Your Neurons: A neural-based approach to answering questions about images'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, pp. 1–9. [Link] (see page 59).

Mania, Horia, Aurelia Guy and Benjamin Recht (2018). *Simple random search provides a competitive approach to reinforcement learning*. arXiv: 1803.07055 (see page 25).

Manning, Christopher D. (Feb. 2011). 'Part-of-speech Tagging from 97% to 100%: Is It Time for Some Linguistics?' In: *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*. Tokyo, Japan, pp. 171–189. [Link] (see page 27, 51).

Mansimov, Elman and Kyunghyun Cho (2018). *Simple Nearest Neighbor Policy Method for Continuous Control Tasks*. [Link] (see page 25).

Mao, Junhua, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan Yuille and Kevin Murphy (June 2016). 'Generation and Comprehension of Unambiguous Object Descriptions'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 11–20. [Link] (see page 28, 31).

Marcus, Gary (2018). *Deep Learning: A Critical Appraisal*. arXiv: 1801.00631 (see page 36, 54).

Marcus, Mitchell P., Mary Ann Marcinkiewicz and Beatrice Santorini (1993). 'Building a Large Annotated Corpus of English: The Penn Treebank'. In: *Computational Linguistics* 19.2, pp. 313–330. [Link] (see page 27).

Marvin, Rebecca and Tal Linzen (Nov. 2018). 'Targeted Syntactic Evaluation of Language Models'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 1192–1202. [Link] (see page 37).

Mascharka, David, Philip Tran, Ryan Soklaski and Arjun Majumdar (June 2018). 'Transparency by Design: Closing the Gap Between Performance and Interpretability in Visual Reasoning'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA, pp. 4942–4950. [Link] (see page 62).

Massiceti, Daniela, Puneet K. Dokania, N. Siddharth and Philip H. S. Torr (2018). *On the State of the Art of Evaluation in Neural Language Models*. arXiv: 1812.06417 (see page 25).

Melis, Gábor, Chris Dyer and Phil Blunsom (2017). *On the State of the Art of Evaluation in Neural Language Models*. arXiv: 1707.05589 (see page 24).

Merity, Stephen, Nitish Shirish Keskar and Richard Socher (2018). *An Analysis of Neural Language Modeling at Multiple Scales*. arXiv: 1803.08240 (see page 24).

Mhasawade, Vishwali, Ildikó Emese Szabó, Melanie Tosik and Sheng-Fu Wang (2018). *Neural Networks and Quantifier Conservativity: Does Data Distribution Affect Learnability?* arXiv: 1809.05733 (see page 39).

Mitchell, Jeff, Pontus Stenetorp, Pasquale Minervini and Sebastian Riedel (June 2018). 'Extrapolation in NLP'. In: *Proceedings of the NAACL Workshop on Generalization in the Age of Deep Learning*. New Orleans, LA, USA, pp. 28–33. [Link] (see page 35, 54).

Moosavi, Nafise Sadat and Michael Strube (Aug. 2017). 'Lexical Features in Coreference Resolution: To be Used With Caution'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada, pp. 14–19. [Link] (see page 22).

Mudrakarta, Pramod Kaushik, Ankur Taly, Mukund Sundararajan and Kedar Dhamdhere (July 2018). 'Did the Model Understand the Question?' In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 1896–1906. [Link] (see page 28, 30, 31, 38, 61).

Naik, Aakanksha, Abhilasha Ravichander, Norman Sadeh, Carolyn Rose and Graham Neubig (Aug. 2018). 'Stress Test Evaluation for Natural Language Inference'. In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*. Santa Fe, NM, USA, pp. 2340–2353. [Link] (see page 38).

Nash, Charlie, Sebastian Nowozin and Nate Kushman (2018). *Generative Entity Networks: Disentangling Entitites and Attributes in Visual Scenes using Partial Natural Language Descriptions*. [Link] (see page 144).

Nematzadeh, Aida, Kaylee Burns, Erin Grant, Alison Gopnik and Tom Griffiths (Nov. 2018). 'Evaluating Theory of Mind in Question Answering'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 2392–2400. [Link] (see page 39, 120).

Nguyen, Anh, Jason Yosinski and Jeff Clune (June 2015). 'Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA, pp. 427–436. [Link] (see page 22).

O'Sullivan, Lewis and Shane Steinert-Threlkeld (June 2019). 'Neural Models of the Psychosemantics of 'Most''. In: *Proceedings of the NAACL Workshop on Cognitive Modeling and Computational Linguistics*. Minneapolis, Minnesota, pp. 140–151. [Link] (see page 137).

Oepen, Stephan and J. Lønning (May 2006). 'Discriminant-Based MRS Banking'. In: *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC)*. Genoa, Italy. [Link] (see page 79).

Packard, Woodley (2018). *ACE: the Answer Constraint Engine*. [Link] (see page 78).

Paperno, Denis (Nov. 2018). 'Limitations in learning an interpreted language with recurrent models'. In: *Proceedings of the EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium, pp. 384–386. [Link] (see page 36).

Papineni, Kishore, Salim Roukos, Todd Ward and Wei-Jing Zhu (July 2002). 'BLEU: A Method for Automatic Evaluation of Machine Translation'. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*. Philadelphia, Pennsylvania, pp. 311–318. [Link] (see page 138).

Perez, Ethan, Florian Strub, Harm de Vries, Vincent Dumoulin and Aaron Courville (Feb. 2018). 'FiLM: Visual Reasoning with a General Conditioning Layer'. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA. [Link] (see page 62, 92, 93, 95, 96, 98, 99, 120, 122, 133).

Petrochuk, Michael and Luke Zettlemoyer (Nov. 2018). 'SimpleQuestions Nearly Solved: A New Upperbound and Baseline Approach'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 554–558. [Link] (see page 24, 28).

Pezzelle, Sandro, Ionut-Teodor Sorodoc and Raffaella Bernardi (June 2018). 'Comparatives, Quantifiers, Proportions: a Multi-Task Model for the Learning of Quantities from Vision'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 419–430. [Link] (see page 137).

Pfungst, Oskar and Carl Leo Rahn (1911). *Clever Hans (the horse of Mr. Von Osten): A Contribution to Experimental Animal and Human Psychology*. Henry Holt and Company. [Link] (see page 42).

Pietroski, Paul, Jeffrey Lidz, Tim Hunter and Justin Halberda (2009). 'The Meaning of 'Most': Semantics, Numerosity and Psychology'. In: *Mind & Language* 24.5, pp. 554–585. [Link] (see page 120, 126, 129–133, 137).

Pinto, Nicolas, David D. Cox and James J. DiCarlo (2008). 'Why is Real-World Visual Object Recognition Hard?' In: *PLOS Computational Biology* 4.1, pp. 1–6. [Link] (see page 26, 41, 51).

Poliak, Adam, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger and Benjamin van Durme (June 2018). 'Hypothesis Only Baselines for Natural Language Inference'. In: *Proceedings of the NAACL Workshop on Lexical and Computational Semantics*. New Orleans, LA, USA, pp. 180–191. [Link] (see page 29).

Ponce, J., T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, Antonio Torralba, C. K. I. Williams, J. Zhang and A. Zisserman (2006). 'Dataset Issues in Object Recognition'. In: *Toward Category-Level Object Recognition*. Ed. by Jean Ponce, Martial Hebert, Cordelia Schmid and Andrew Zisserman. Vol. 4170. LNCS. Springer Science+Business Media, pp. 29–48. [Link] (see page 21, 23, 27).

Post, Matt (Nov. 2018). 'A Call for Clarity in Reporting BLEU Scores'. In: *Proceedings of the 3rd Conference on Machine Translation*. Belgium, Brussels, pp. 186–191. [Link] (see page 32).

Press, Ofir and Noah A. Smith (2018). *You May Not Need Attention*. arXiv: 1810.13409 (see page 24).

Rajeswaran, Aravind, Kendall Lowrey, Emanuel V. Todorov and Sham M. Kakade (Dec. 2017). 'Towards Generalization and Simplicity in Continuous Control'. In: *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS)*. Long Beach, CA, USA, pp. 6550–6561. [Link] (see page 25).

Rajpurkar, Pranav, Robin Jia and Percy Liang (July 2018). 'Know What You Don't Know: Unanswerable Questions for SQuAD'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 784–789. [Link] (see page 34).

Rajpurkar, Pranav, Jian Zhang, Konstantin Lopyrev and Percy Liang (Nov. 2016). 'SQuAD: 100,000+ Questions for Machine Comprehension of Text'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Austin, TX, USA, pp. 2383–2392. [Link] (see page 29, 34, 121).

Raposo, David, Adam Santoro, David Barrett, Razvan Pascanu, Timothy Lillicrap and Peter W. Battaglia (2017). *Discovering objects and their relations from entangled scene representations*. arXiv: 1702.05068 (see page 62, 63).

Recht, Benjamin, Rebecca Roelofs, Ludwig Schmidt and Vaishaal Shankar (2018). *Do CIFAR-10 Classifiers Generalize to CIFAR-10?* arXiv: 1806.00451 (see page 23).

— (2019). *Do ImageNet Classifiers Generalize to ImageNet?* arXiv: 1902.10811 (see page 23).

Reimers, Nils and Iryna Gurevych (2018). *Why Comparing Single Performance Scores Does Not Allow to Draw Conclusions About Machine Learning Approaches*. arXiv: 1803.09578 (see page 33, 40).

Reiter, Ehud (2018). 'A Structured Review of the Validity of BLEU'. In: *Computational Linguistics* 44.3. [Link] (see page 32).

Ren, Mengye, Ryan Kiros and Richard S. Zemel (Dec. 2015). 'Exploring Models and Data for Image Question Answering'. In: *Proceedings of the 29th International Conference on Neural Information Processing Systems (NIPS)*. Montreal, Canada, pp. 2953–2961. [Link] (see page 58, 59, 61).

Rimell, Laura and Stephen Clark (Oct. 2008). 'Adapting a Lexicalized-grammar Parser to Contrasting Domains'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Honolulu, HI, USA, pp. 475–484. [Link] (see page 27).

Ritter, Samuel, David Barrett, Adam Santoro and Matt M. Botvinick (Aug. 2017). 'Cognitive Psychology for Deep Neural Networks: A Shape Bias Case Study'. In: *Proceedings of the*

34th International Conference on Machine Learning (ICML). Sydney, Australia, pp. 2940–2949. [Link] (see page 39, 120, 137).

Rosenfeld, Amir, Richard S. Zemel and John K. Tsotsos (2018). *The Elephant in the Room*. arXiv: 1808.03305 (see page 37, 118).

Santoro, Adam, David Raposo, David Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia and Timothy Lillicrap (Dec. 2017). 'A simple neural network module for relational reasoning'. In: *Proceedings of the 31th International Conference on Neural Information Processing Systems (NIPS)*. Long Beach, CA, USA, pp. 4967–4976. [Link] (see page 25, 62, 92–94, 98, 100, 122, 137).

Scheffler, Konrad and Steve Young (June 2001). 'Corpus-based dialogue simulation for automatic strategy learning and evaluation'. In: *Proceedings of the NAACL Workshop on Adaptation in Dialogue Systems*. Pittsburgh, PA, USA, pp. 64–70. [Link] (see page 35).

Scheffler, Konrad and Steve Young (Mar. 2002). 'Automatic Learning of Dialogue Strategy Using Dialogue Simulation and Reinforcement Learning'. In: *Proceedings of the 2n International Conference on Human Language Technology Research (HLT)*. San Diego, CA, USA, pp. 12–19. [Link] (see page 35).

Schneider, Nathan, Vivek Srikumar, Jena D. Hwang and Martha Palmer (June 2015). 'A Hierarchy with, of, and for Preposition Supersenses'. In: *Proceedings of the NAACL Workshop on Linguistic Annotation*. Denver, CO, USA, pp. 112–123. [Link] (see page 27).

Sculley, D., Jasper Snoek, Alex Wiltschko and Ali Rahimi (May 2018). 'Winner's Curse? On Pace, Progress, and Empirical Rigor'. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada. [Link] (see page 20, 40, 41).

Shekhar, Ravi, Sandro Pezzelle, Yauhen Klimovich, Aurélie Herbelot, Moin Nabi, Enver Sangineto and Raffaella Bernardi (Aug. 2017). 'FOIL it! Find One mismatch between Image and Language caption'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada, pp. 255–265. [Link] (see page 30, 37).

Shen, Dinghan, Guoyin Wang, Wenlin Wang, Martin Renqiang Min, Qinliang Su, Yizhe Zhang, Chunyuan Li, Ricardo Henao and Lawrence Carin (July 2018). 'Baseline Needs More Love: On Simple Word-Embedding-Based Models and Associated Pooling Mechanisms'. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*. Melbourne, Australia, pp. 440–450. [Link] (see page 25).

Shih, Kevin J., Saurabh Singh and Derek Hoiem (June 2016). 'Where to Look: Focus Regions for Visual Question Answering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 4613–4621. [Link] (see page 59).

Shwartz-Ziv, Ravid and Naftali Tishby (2017). *Opening the Black Box of Deep Neural Networks via Information*. arXiv: 1703.00810 (see page 56).

Siegel, Melanie, Emily M. Bender and Francis Bond (2016). *Jacy: An Implemented Grammar of Japanese*. Studies in Computational Linguistics. CSLI Publications. [Link] (see page 87).

Smith, Noah A. (2012). *Adversarial Evaluation for Models of Natural Language*. arXiv: 1207.0245 (see page 28, 32, 35, 51).

Sorodoc, Ionut-Teodor, Sandro Pezzelle, Aurélie Herbelot, M. Dimiccoli and Raffaella Bernardi (2018). 'Learning quantification from images: A structured neural architecture'. In: *Natural Language Engineering* 24.3, pp. 363–392. [Link] (see page 137).

Sorodoc, Ionut, Angeliki Lazaridou, Gemma Boleda, Aurélie Herbelot, Sandro Pezzelle and Raffaella Bernardi (Aug. 2016). '"Look, some Green Circles!": Learning to Quantify from Images'. In: *Proceedings of the ACL Workshop on Vision and Language*. Berlin, Germany, pp. 75–79. [Link] (see page 137).

Sproat, Richard and Navdeep Jaitly (2016). *RNN Approaches to Text Normalization: A Challenge*. arXiv: 1611.00068 (see page 21, 22, 30).

Srikumar, Vivek and Dan Roth (2013). 'Modeling Semantic Relations Expressed by Prepositions'. In: *Transactions of the Association for Computational Linguistics* 1, pp. 231–242. [Link] (see page 27).

Srinivasan, Siddarth, Richa Arora and Mark Riedl (June 2018). 'A Simple and Effective Approach to the Story Cloze Test'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 92–96. [Link] (see page 24, 31).

Stanovsky, Gabriel and Mark Hopkins (Nov. 2018). 'Spot the Odd Man Out: Exploring the Associative Power of Lexical Resources'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 1533–1542. [Link] (see page 37).

Stoianov, Ivilin and Marco Zorzi (2012). 'Emergence of a 'visual number sense' in hierarchical generative models'. In: *Nature Neuroscience* 15.2, pp. 194–196. [Link] (see page 136).

Strubell, Emma, Ananya Ganesh and Andrew McCallum (July 2019). 'Energy and Policy Considerations for Deep Learning in NLP'. In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*. Florence, Italy. [Link] (see page 56).

Sturm, Bob L. (2014). 'A Simple Method to Determine if a Music Information Retrieval System is a "Horse"'. In: *IEEE Transactions on Multimedia* 16.6, pp. 1636–1644. [Link] (see page 23, 38, 41, 42).

Suarez, Joseph, Justin Johnson and Fei-Fei Li (2018). *DDRprog: A CLEVR Differentiable Dynamic Reasoning Programmer*. arXiv: 1803.11361 (see page 62).

Suhr, Alane, Mike Lewis, James Yeh and Yoav Artzi (Aug. 2017). 'A Corpus of Natural Language for Visual Reasoning'. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vancouver, Canada, pp. 217–223. [Link] (see page 34, 62).

Sulem, Elior, Omri Abend and Ari Rappoport (Nov. 2018). 'BLEU is Not Suitable for the Evaluation of Text Simplification'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 738–744. [Link] (see page 32).

Sun, Chen, Abhinav Shrivastava, Saurabh Singh and Abhinav Gupta (Oct. 2017). 'Revisiting Unreasonable Effectiveness of Data in Deep Learning Era'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Venice, Italy, pp. 843–852. [Link] (see page 20).

Suzgun, Mirac, Yonatan Belinkov and Stuart M. Shieber (Jan. 2019). 'On Evaluating the Generalization of LSTM Models in Formal Languages'. In: *Proceedings of the Society for Computation in Linguistics (SCiL)*. New York, NY, USA, pp. 277–286. [Link] (see page 35).

Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens and Zbigniew Wojna (June 2016). 'Rethinking the Inception Architecture for Computer Vision'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 2818–2826. [Link] (see page 141).

Szegedy, Christian, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow and Rob Fergus (Apr. 2014). 'Intriguing properties of neural networks'. In: *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*. Banff, Canada. [Link] (see page 22).

Szucs, Denes and John P. A. Ioannidis (2017). 'When Null Hypothesis Significance Testing Is Unsuitable for Research: A Reassessment'. In: *Frontiers in Human Neuroscience* 11. [Link] (see page 33, 40).

Talman, Aarne and Stergios Chatzikyriakidis (Aug. 2019). 'Testing the Generalization Power of Neural Network Models Across NLI Benchmarks'. In: *Proceedings of the ACL Workshop on BlackboxNLP*. Florence, Italy. [Link] (see page 22).

Tapaswi, Makarand, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun and Sanja Fidler (June 2016). 'MovieQA: Understanding Stories in Movies through Question-Answering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 4631–4640. [Link] (see page 67).

Teney, Damien, Peter Anderson, Xiaodong He and Anton van den Hengel (June 2018). 'Tips and Tricks for Visual Question Answering: Learnings From the 2017 Challenge'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA, pp. 4223–4232. [Link] (see page 59).

Theis, L., A. van den Oord and Matthias Bethge (May 2016). 'A note on the evaluation of generative models'. In: *Proceedings of the 4th International Conference on Learning Representations (ICLR)*. San Juan, Puerto Rico. [Link] (see page 32).

Thomason, Jesse, Daniel Gordon and Yonatan Bisk (June 2019). 'Shifting the Baseline: Single Modality Performance on Visual Navigation & QA'. In: *Proceedings of the Conference of*

the North American Chapter of the Association for Computational Linguistics (NAACL). Minneapolis, MN, USA, pp. 1977–1983. [Link] (see page 28).

Tobin, Josh, Rachel Fong, Alex Ray, Jonas Schneider, Wojciech Zaremba and Pieter Abbeel (Sept. 2017). 'Domain randomization for transferring deep neural networks from simulation to the real world'. In: *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. Vancouver, Canada, pp. 23–30. [Link] (see page 147).

Tommasi, Tatiana, Novi Patricia, Barbara Caputo and Tinne Tuytelaars (Oct. 2015). 'A Deeper Look at Dataset Bias'. In: *Proceedings of the 37th German Conference on Pattern Recognition (GCPR)*. Aachen, Germany, pp. 504–516. [Link] (see page 26, 27, 51).

Torralba, Antonio and A. A. Efros (June 2011). 'Unbiased Look at Dataset Bias'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado Springs, CO, USA, pp. 1521–1528. [Link] (see page 26, 27, 35, 41, 42, 55).

Trichelair, Paul, Ali Emami, Jackie Chi Kit Cheung, Adam Trischler, Kaheer Suleman and Fernando Diaz (2018). *A Simple Method for Commonsense Reasoning*. arXiv: 1811.01778 (see page 30, 38).

Trinh, Trieu H. and Quoc V. Le (2018). *A Simple Method for Commonsense Reasoning*. arXiv: 1806.02847 (see page 24).

Ulyanov, D., A. Vedaldi and V. Lempitsky (June 2018). 'Deep Image Prior'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Salt Lake City, UT, USA. [Link] (see page 24).

Vedantam, Ramakrishna, C. Lawrence Zitnick and Devi Parikh (June 2015). 'CIDEr: Consensus-based image description evaluation'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA, pp. 4566–4575. [Link] (see page 138).

Vinyals, Oriol, Alexander Toshev, Samy Bengio and Dumitru Erhan (June 2015). 'Show and tell: A neural image caption generator'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, MA, USA, pp. 3156–3164. [Link] (see page 141).

Wagstaff, Kiri (June 2012). 'Machine Learning that Matters'. In: *Proceedings of the 29th International Conference on Machine Learning (ICML)*. Edinburgh, United Kingdom. [Link] (see page 22, 41).

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy and Samuel R. Bowman (Nov. 2018a). 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. In: *Proceedings of the EMNLP Workshop on BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium, pp. 353–355. [Link] (see page 27, 35).

Wang, Josiah, Pranava Madhyastha and Lucia Specia (June 2018b). 'Object Counts! Bringing Explicit Detections Back into Image Captioning'. In: *Proceedings of the Conference of the*

North American Chapter of the Association for Computational Linguistics (NAACL). New Orleans, LA, USA, pp. 2180–2193. [Link] (see page 24, 29).

Weber, Noah, Leena Shekhar and Niranjan Balasubramanian (June 2018). 'The Fine Line between Linguistic Generalization and Failure in Seq2Seq-Attention Models'. In: *Proceedings of the NAACL Workshop on Generalization in the Age of Deep Learning*. New Orleans, LA, USA, pp. 24–27. [Link] (see page 36).

Weston, Jason, Antoine Bordes, Sumit Chopra and Tomas Mikolov (2015). *Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks*. arXiv: 1502.05698 (see page 30, 36, 61, 62, 80, 141).

Wieting, John and Douwe Kiela (2019). *No Training Required: Exploring Random Encoders for Sentence Classification*. arXiv: 1901.10444 (see page 25).

Williams, Adina, Andrew Drozdov and Samuel R. Bowman (2018a). 'Do latent tree learning models identify meaningful structure in sentences?' In: *Transactions of the Association for Computational Linguistics* 6, pp. 253–267. [Link] (see page 21).

Williams, Adina, Nikita Nangia and Samuel R. Bowman (June 2018b). 'A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference'. In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. New Orleans, LA, USA, pp. 1112–1122. [Link] (see page 23, 29, 34).

Wu, Xiaolin, Xi Zhang and Xiao Shu (2018). *On Numerosity of Deep Convolutional Neural Networks*. arXiv: 1802.05160 (see page 36, 137).

Xie, Huiyuan, Tom Sherborne, Alexander Kuhnle and Ann Copestake (Feb. 2020). 'Going beneath the surface: Evaluating image captioning for grammaticality, truthfulness and diversity'. In: *Proceedings of the AAAI Workshop on Evaluating Evaluation of AI Systems*. New York, NY, USA. [Link] (see page 17, 138).

Xiong, Caiming, Stephen Merity and Richard Socher (June 2016). 'Dynamic Memory Networks for Visual and Textual Question Answering'. In: *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML)*. New York, NY, USA, pp. 2397–2406. [Link] (see page 59).

Xu, Huijuan and Kate Saenko (Oct. 2016). 'Ask, Attend and Answer: Exploring Question-Guided Spatial Attention for Visual Question Answering'. In: *Proceedings of the 14th European Conference on Computer Vision (ECCV)*. Amsterdam, The Netherlands, pp. 451–466. [Link] (see page 59).

Xu, Qiantong, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu and Kilian Weinberger (2018). *An empirical study on evaluation metrics of generative adversarial networks*. arXiv: 1806.07755 (see page 32).

Yang, Robert Guangyu, Igor Ganichev, Xiao Jing Wang, Jonathon Shlens and David Sussillo (Sept. 2018). 'A Dataset and Architecture for Visual Reasoning with a Working Memory'.

In: *Proceedings of the 15th European Conference on Computer Vision (ECCV)*. Munich, Germany, pp. 729–745. [Link] (see page 62, 70, 136).

Yang, Zichao, Xiaodong He, Jianfeng Gao, Li Deng and Alexander J. Smola (June 2016). 'Stacked Attention Networks for Image Question Answering'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 21–29. [Link] (see page 59, 92, 94).

Yu, Licheng, Eunbyung Park, Alexander C. Berg and Tamara L. Berg (Dec. 2015). 'Visual Madlibs: Fill in the Blank Description Generation and Question Answering'. In: *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. Santiago, Chile, pp. 2461–2469. [Link] (see page 58).

Yuille, Alan and Chenxi Liu (2018). 'Deep Nets: What have they ever done for Vision?' In: *CBMM Memos* 088. [Link] (see page 23).

Zellers, Rowan, Yonatan Bisk, Roy Schwartz and Yejin Choi (Nov. 2018). 'SWAG: A Large-Scale Adversarial Dataset for Grounded Commonsense Inference'. In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Brussels, Belgium, pp. 93–104. [Link] (see page 35).

Zhang, Chiyuan, Samy Bengio, Moritz Hardt, Benjamin Recht and Oriol Vinyals (Apr. 2017a). 'Understanding deep learning requires rethinking generalization'. In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France. [Link] (see page 22).

Zhang, Chiyuan, Oriol Vinyals, Rémi Munos and Samy Bengio (2018a). *A Study on Overfitting in Deep Reinforcement Learning*. arXiv: 1804.06893 (see page 36).

Zhang, Jianming, Shugao Ma, Mehrnoosh Sameki, Stan Sclaroff, Margrit Betke, Zhe Lin, Xiaohui Shen, Brian Price and Radomír Měch (2017b). 'Salient Object Subitizing'. In: *International Journal of Computer Vision* 124.2, pp. 169–186. [Link] (see page 136).

Zhang, Peng, Yash Goyal, Douglas Summers-Stay, Dhruv Batra and Devi Parikh (June 2016). 'Yin and Yang: Balancing and Answering Binary Visual Questions'. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Las Vegas, NV, USA, pp. 5014–5022. [Link] (see page 28, 34, 37, 60, 61).

Zhang, Quanshi, Wenguan Wang and Song-Chun Zhu (Feb. 2018b). 'Examining CNN Representations With Respect to Dataset Bias'. In: *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*. New Orleans, LA, USA, pp. 4464–4473. [Link] (see page 21).

Zhang, Yan, Jonathon Hare and Adam Prügel-Bennett (May 2018c). 'Learning to Count Objects in Natural Images for Visual Question Answering'. In: *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. Vancouver, Canada. [Link] (see page 137).

Zhu, Yuke, Oliver Groth, Michael Bernstein and Li Fei-Fei (June 2016). 'Visual7W: Grounded Question Answering in Images'. In: *Proceedings of the IEEE Conference on Computer*

Vision and Pattern Recognition (CVPR). Las Vegas, NV, USA, pp. 4995–5004. [Link] (see page 57–59).

Zitnick, C. Lawrence, Ramakrishna Vedantam and Devi Parikh (2016). 'Adopting Abstract Images for Semantic Scene Understanding'. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38.4, pp. 627–638. [Link] (see page 36, 80).